What is Statistics?

Statistics is the study of how to collect, organize, and interpret numerical information and data.
Statistics is both the science of uncertainty and the technology of extracting information from data.
Statistics is used to help to us make decisions. This is especially important in health care and public health.

Example: CDC & the Flu Vaccine

During the year, the United States Center for Disease Control and Prevention (CDC) collects, organizes, analyzes and interprets numerical information and data
They extract information from the data and make decisions about what to include in next year's flu vaccine.

Individuals & Variables

Meaning Outside Statistics	Meaning in Statistics
Individuals are people We expect 50 individuals at the graduation.	Individuals are people or objects include in a study. 5 Individuals could be 5 people, 5 records, or 5 reports.
A variable is a factor that can vary possibly causing a problem The time the shop takes with my car is an unknown variable, and I can't predict it.	A variable is a characteristic of the individual to be measured or observed. The age of Individual person The time an individual record was entered The diagnosis listed on an individual report

Concepts in Statistics

Statistics is used in healthcare and other disciplines to help aid in decision-making.
Understanding Statistics is necessary to understand certain processes in healthcare.

What is a Population?

Definition	Example
A population is a group of people or objects with a common theme.	Theme: Nurse who work at (MGH)
When every member of that group is considered, it is a population. In population data, data from every individual in the population in the population in the population is available. Entire population = census	Population: List from Human Resources of every currently employed nurse at MGH

What is a Sample?

Definition	Example
A sample is a small portion of the population. It can be a representative sample.	Only survey ICU nurses at MGH Not a representative sample
But it can also be a biased sample. In sample data, data is only available from some of the individuals in the population Very commonly used in research studies of patients	At least one nurse from each department More representative sample

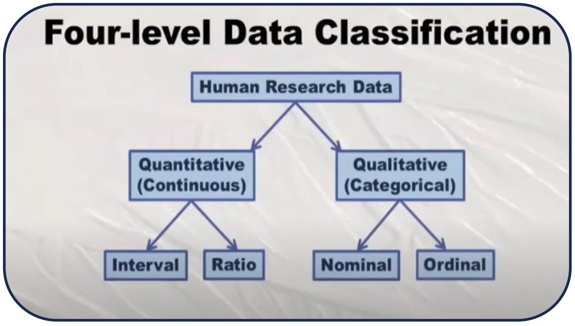
Note: Total Population = N
Total Sample of Population = n

Parameters vs. Statistics

Parameters	**Statistics**
A Parameter is a measure that describes the entire population	A Statistics is a measure that describes only a sample of a population

Describing vs. Inferring

Descriptive statistics	Inferential statistics
Descriptive statistics involve methods of organizing, picturing, and summarizing information from samples and populations.	Inferential statistics involves methods of using information from a sample to draw conclusions regarding the population.


Four-level Data Classification
Human Research Data
Quantitative (Continuous)
Qualitative (Categorical)
Interval
Ratio
Nominal
Ordinal

Classifying Levels of Measurement (Four -level system**)**
Classifying Variables

Quantitative (Continuous) vs. Qualitative (Categorical) data

Interval vs. ratio data

Nominal vs. ordinal data

Quantitative is a numerical measurement of something.

Qualitative refers to a "quality" or categorical characteristic of something

Simulation

A simulation is defined as a "numerical facsimile or representation of a real- world phenomenon."

It is an essentially working through a pretend situation to see how it would come out in the case it was real.

That is why this course includes many simulations, or real- life examples.

Concepts in Sampling

It is important to do your best to avoid non- sampling error

This is achieved by making sure you do not have under coverage when sampling from your sampling frame.

Definition & Example

Definition	**Example**
	You have a list of the population of students in a class.
	You want to take a sample of 5 (n=5)
"A simple random sample of n measurements from a population is a subset of the population is a subset of the population selected in such a manner that every sample of size n from the population has an equal chance of being selected."	If you take a sample random sample (SRS) from the class list, it means all the different possible groups of 5 students you could pick from the list has an equal chance of being the sample (group) you actually pick.

One Method of SRS

Number all of the individuals in the population with a unique number.

Like student ID number

Put all the student ID number in a place from which you can draw randomly with looking (like a hat)

Draw 5 ID's and use those students as your sample.

Another Method of SRS

Generate a list of random numbers as long as the list of the population.

Randomly assign these numbers of the population in the list.

Take the first 5 numbers (whoever gets assigned 1 through 5).

SRS Means equal chance of being selected

First Method: old- fashioned "hat"

Second Method: Electronic "hat"

In Both Method, all members of the population had an equal population had an equal probability of being selected into the sample

Limits of simple Random Sampling (SRS)

1. You need a list
   you don't know who will present at the Emergency Department that day, how do you sample?
   Okay when a list is available.
2. You need a good list
   Otherwise, you risk under coverage
   What if part- time students were not on the list?
   Non- sampling error

What is Stratified Sampling?

First, the list is divided into group, or strata.
This is a way to make it so that there are certain proportions of groups in the final sample.
Next, sampling (SRS) takes place for each of the strata.

Steps in Stratified Sampling

1. Divide entire population into distinct subgroups called strata.
2. The strata are based on a specific characteristic, such as age, income, education level, and so on.
3. All members of a stratum share this specific characteristic.
4. Draw an SRS from each stratum.

Limitations of stratified Sampling

1. Oversampling one group means your summary statistic is unbalanced
2. It is not possible to do without a list beforehand (life with SRS)
3. It also is hard because you have to spit the list into groups(“STRATA”) then SRS from the strata.

Systematic Sampling

Systematic Sampling can be done with or without a list!
Systematic Sampling is best described though the steps one takes to do it

Step in Systematic Sampling

1. Arrange all individuals of the population in a particular order.
2. Pack a random individual as a start.
3. Then take every kth member of the population in the sample.

Note: - “kth” means “every so many”.

Characteristics of Systematic Sampling

You cannot do this when there is a patter to the data (boy/girl/boy/girl)
You can do Systematic Sampling in a clinical setting in a clinical setting, where you do not know who is going to come in that day

Cluster Sampling

when we have a problem is in a particular geographic lactation.
Cluster sampling is used when geography is important in sampling.
The map is divided into areas, and all the people in a particular area are sampled.
Biased toward type of people living in the area.

Why use Cluster Sampling?

The problem is localized to a particular is location.
In cluster sampling, we begin by dividing the map in geographic areas.
Then we randomly pick clusters, or areas, from the map. We take all the people in the cluster.

Problems with Cluster Sampling

Sometimes, the people located in a cluster are all similar in a way that makes the problem hard to study.
If cancer rates are high all over the clusters, it’s hard to see if a geographic location is causing higher rates.

Convenience Sampling

Convenience Sampling can be used under low-risk circumstances. However, often results are not reliable

What is Convenience Sampling?

Using results or data that are conveniently or readily obtained.
Can be useful if not a lot of resources allocated to the study.
Use an already- assembled group for surveys.
Ask patients in the waiting room to fill out a survey, or students in a class.

What are the Problems with Convenience Sapling?

There is a bias in every group.

Often miss important subpopulations (what stratified sampling addresses).

Results can be severely biased

Multi- stage Sampling

Combination of sampling strategies layered in stages.

Example:

1. Stage 1: Cluster sample of states (two census regions)
2. Stage 2: Simple random sample of counties (from each state)
3. Stage 3: Stratified sample of schools (urban/rural)
4. Stage 4: Stratified sample of classrooms

Convenience & Multi- Stage Sampling

Avoid using convenience sampling unless the question is low risk

Use if the only type of sampling possible under the circumstances

Also used when resources are low

Multi- stage sampling usually used in large, governmental studies.

Basic Guidelines for planning a statistical study

1. State a hypothesis.
2. Identify the individuals of interest.
3. Specify the variables to measure.
4. Determine if you will use the entire population or a sample.
   If you choose a sample, choose a sampling method
5. Address ethical concerns before data collection.
6. Collect the data.
7. Use descriptive or inferential statistics your hypothesis.
8. Note any concerns about your data collection or analysis
   Make recommendations for future studies

Hypothesis & Variables (first three point of Basic Guidelines)

Hypothesis: Air pollution causes asthma in children who live in urban settings

Individuals: Children in urban settings

Variables: Air pollution and asthma

Sampling, Ethics & Data Collection

1. Either collect data or use existing dataset
   - Can use a government dataset for population measures
2. Can collect data from a sample for estimates
   - Need to choose sampling approach
   - Will need consent if legally found to be “human research”
   - May need consent from parents to collect data about children

Census vs. Sample

In a census, measurements or observations from the entire population are used.	In a sample, measurements or observations from part of the population are used.

Experiment vs. Observational Study

Experiment Study	**Observational Study**
A treatment or intervention is deliberately assigned on the individuals	Observations and measurements of individuals are taken
The purpose is to study the possible effect of the treatment or intervention on the variables measured	However, no treatment or intervention is assigned by the researcher

Replication

Studies must be done rigorously enough to be replicated.

Replicating the results of observational studies and experiments is necessary for science to progress.

Bias

Surveys can provide a lot of useful information However; it is important that all aspects of survey design and administration minimize "bias".

Several considerations should be made

Non-response & Voluntary Response

If many people refuse your survey, the people who do complete it are likely to have a biased opinion.

There may be a reason they do not complete your survey that has to do with how they feel about your survey topic.

Truthfulness of Response

- Respondents may lie on purpose
  - If asked a question that is too personal
  - If asked a question too hard to think about
- Respondents may lie inadvertently
  - May not remember if asking about something that happened a long time ago
  - May have "recall bias" influenced by events that have happened since original event

Why Randomize?

Randomization is used to assign individuals to treatment groups. This helps prevent bias in selecting members for each group. It distributes "lurking variables" evenly

Placebo & Placebo Effect

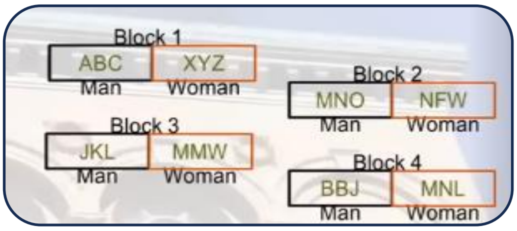
Placebo effect occurs when there is no treatment, but participant assumes s/he is receiving treatment and responds favorably.

The placebo is given to a control group, which receives the placebo (or attention control if treatment is not a drug).

Used as a control or comparison group.

Blocked Randomization

1. If you want men and women equal in two randomized groups, create "blocks" with two slots — one for a man, and one for a woman.
2. As people come in and enroll in the study and you measure them, assign them to blocks.
3. Then, randomize the blocks. You will get a paired couples in each group!



Blinding

Blinding is where a person (participant, research staff) is deliberately not told of a treatment assignment in a study so s/he is not biased in reporting study information.

- Example: A participant is blinded to treatment or placebo.

Double-blind means study staff and participant do not know treatment assignment.

Randomization & Bias

Randomization is used to reduce bias in an experiment.

Blocked randomization can even out groups.

Blinding further prevents bias

The placebo effect is necessary to take into account.

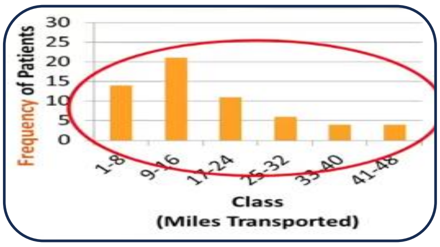
What is a Frequency Histogram?

It's a specific type of bar chart made from data in a frequency table.

Frequency histograms and relative frequency histograms.

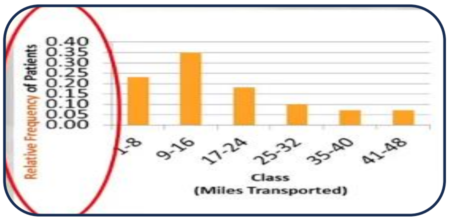
The purpose of the chart is to identify the "distribution" of the data.

Steps to Follow to Draw a Frequency Histogram



1. Make a frequency table.
2. Draw a vertical line for the y-
3. Write "Frequency of along the y-axis.
4. Draw a horizontal line for the x-axis.
5. Write the classes below the x- axis and label them.
6. For the first class, find the frequency in the table. Look for it on the y-axis and draw a horizontal line.
7. Draw two vertical lines down to make a bar.
8. Repeat for all the other classes.
9. Color in the bars

Relative Frequency Histogram



In the relative frequency histogram, the relative frequency goes on the y- axis.
The chart looks takes on a similar pattern.
Relative frequency better for comparing two populations or two samples.

Frequency & Relative Frequency Histograms

After making a frequency table, it is important to also make a frequency histogram and/or a relative frequency histogram.
These are used to reveal the "distribution" in the data

What is a Distribution?

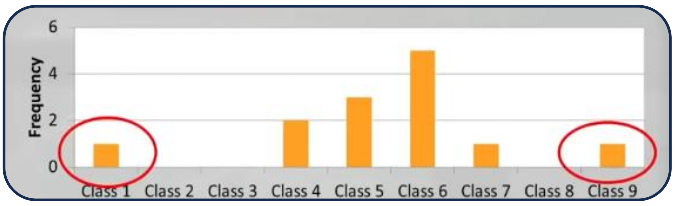
- It is the shape that is made if you draw a line along the edges of a histogram's bars.
- A stem-and-leaf of the same data will make the same shape on its side.

5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution

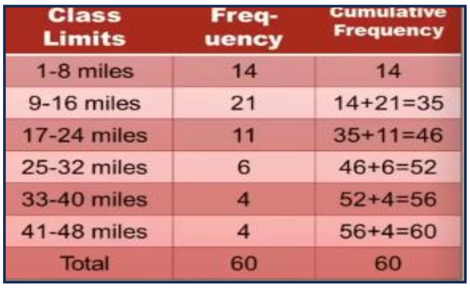
Outliers

Outliers are data values that are "very different" from other measurements in the dataset.



Cumulative Frequency

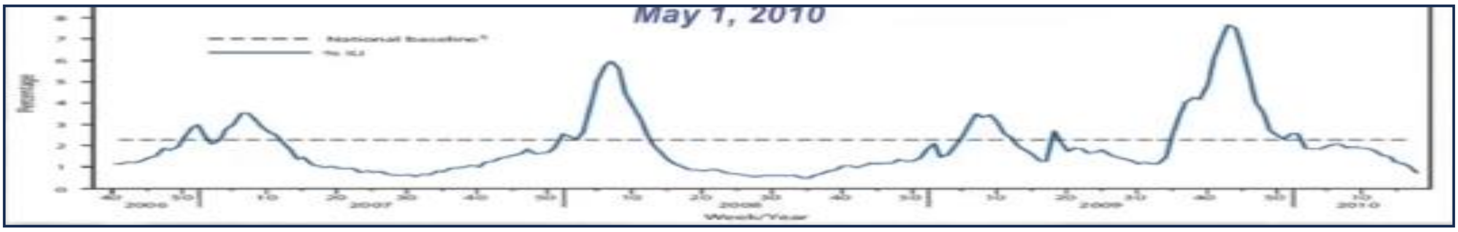
- In "cumulative frequency", you add up all the classes before the class you are on. The first class is always the same as the frequency. Each cumulative frequency is equal to or higher than the last one.



Class Limits	Freq- uency	Cumulative Frequency
1-8 miles	14	14
9-16 miles	21	14+21=35
17-24 miles	11	35+11=46
25-32 miles	6	46+6=52
33-40 miles	4	52+4=56
41-48 miles	4	56+4=60
Total	60	60

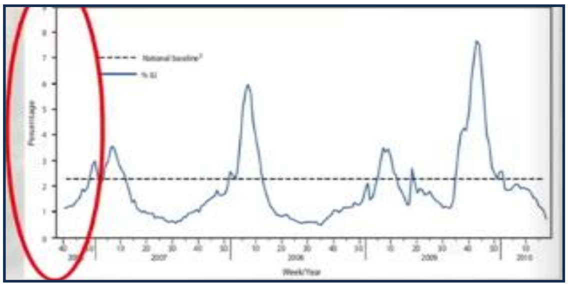
Time Series Graph

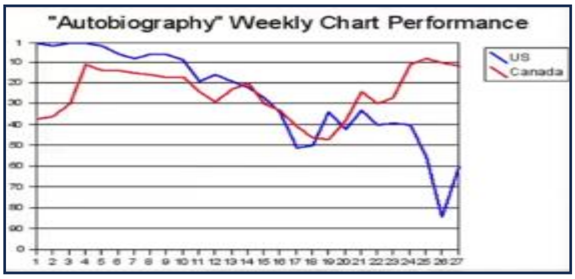
Time Series data are made of measurement for the same variable for the same individual taken at intervals over a period of time. Stock market prices Yearly rates of diseases such as influenza



How to Plot a Time-Series Graph

1. Get time series data: one measurement per time period (month, year, etc.).
2. Draw a horizontal line for the x-axis. Label the time periods
3. Draw a vertical line for the y- axis.
   Make sure it is tall enough for the highest data value. Label it.

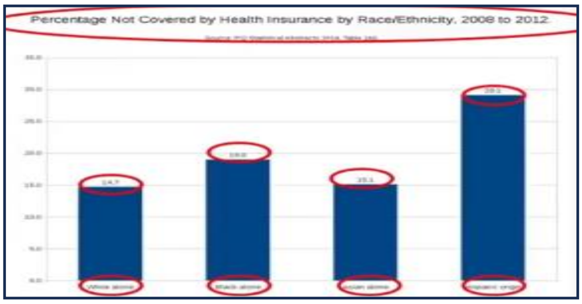




4. Looking at your data, the x- axis, and the y-axis, put dots in where the data points are.
5. Connect the dots.
6. You can have more than one line on the graph for more than one set of data values, but this requires a legend.

Features of a Bar Graph

1. Bars can be vertical or horizontal
2. Are of uniform width and of uniform spacing
3. Length of bars represent variable's frequency or percentage of occurrence.
4. same measurement scale used for each bar.
5. Includes title, bar labels, and scale labels on axis or actual values for each bar.

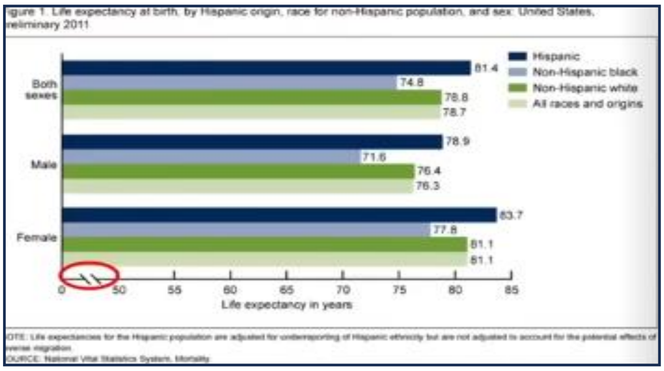


Bar Graph vs. Histogram: What's the Difference?

- The frequency histogram and relative frequency histogram are “Special case” of a bar graph
- They are bar graphs that:
  Must have classes of a quantitative variable on the x-axis
  Must have frequency or relative frequency on the y-axis

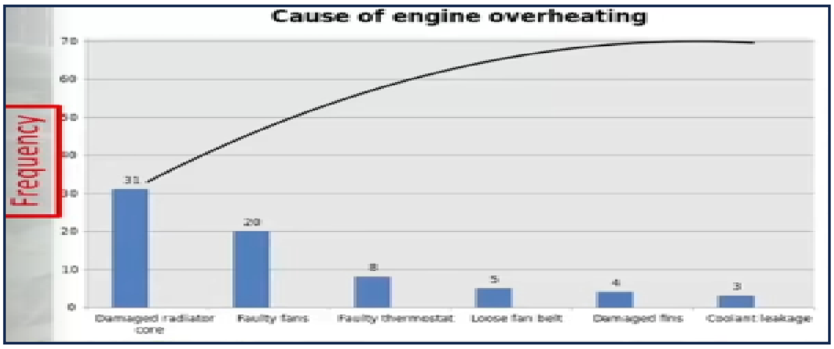
Warning About Changing the Scale

- With a taller y-axis, the differences between bars look less dramatic.
- Clustered means more than one bar is graphed for each category (see legend).
- Also, look for the beginning of the scale. Some do not start at zero, and then the bars do not start at zero.
- Look for the squiggle.



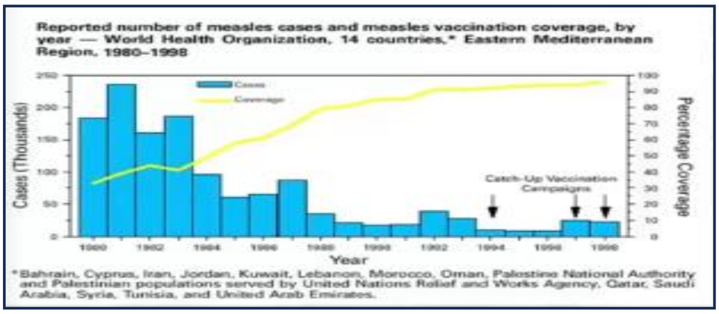
Pareto Chart

- The height of the bar indicates the frequency of an event.
- Arranged left to right according to decreasing height.
- Meant to graph frequencies of "problems"
- Used more in engineering than in healthcare.



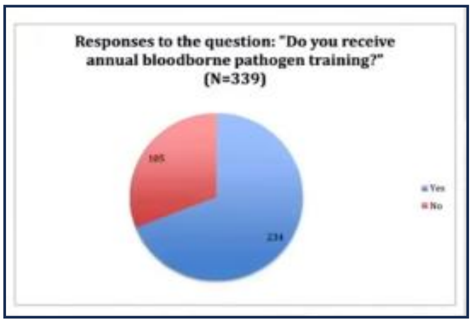
Bar Graph Summary

- Bar graphs must be made following a few rules.
- Can be very helpful for visualizing and comparing quantitative and qualitative data.
- Scales are important in bar graphs!
- Pareto charts not used much in healthcare.



Pie Charts

- Pie chart (also circle graph) used with counts of “mutually exclusive” frequencies
- Often made in graphing programs because difficult to do by hand
- Very common in healthcare



Features of a Pie Chart

- Every individual must be put in only one category.
- Can be qualitative or quantitative variable.
- If quantitative, put in classes and then graphed.

A Few Notes on Pie Charts

- Must be mutually exclusive categories
    - "Favorite" color vs. "check the colors you like"
- More informative to put % than frequency, but it is helpful to do both.
- Always include title and legend.

On All Graphs

- Provide a title
- Label axes
- Identify units of measure
- Make the graph as clear as possible (think of font size, number of items graphed).

Choosing the Right Kind of Graph

Type of Graph	Cases Where Graph is Useful
Frequency Histogram	For quantitative data, when you want to see the distribution.
Relative Frequency Histogram	For quantitative data, when you want to see the distribution. Also, good for comparing to other data.
Stem-and-leaf Display	For quantitative data, when you want to see the distribution. Easier to make by hand than histogram.
Time series graph	For graphing a variable that changes over time and is measured at regular intervals.
Bar graph	For qualitative or quantitative data, and for displaying frequency or percentage.
Pareto chart	For frequencies of rare events in descending order.
Pie Graph	For mutually-exclusive categories (quantitative or qualitative).

Remember Quantitative Data?

1. Qualitative data are categorical
    - Gender, race, diagnosis
2. Quantitative data are numerical
    - Age, heart rate, blood pressure

A Few Definitions

1. Class: An interval in the data.
    - Example: Between 30 and 40 miles.
2. Class limit: The lowest and highest value that can fit in a class.
    - Example: 30 would be the lower-class limit, and 40 would be the upper-class limit.
3. Class width: How wide the class is.
    - Example: Upper class limit (40) minus lower class limit (30) = 10, then add 1 = 11.
    - Example: 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 = 11 numbers
4. Frequency: How many values from the data fall in the class.
    - Example: How many patients were transported 30 to 40 miles.

Decide on Classes

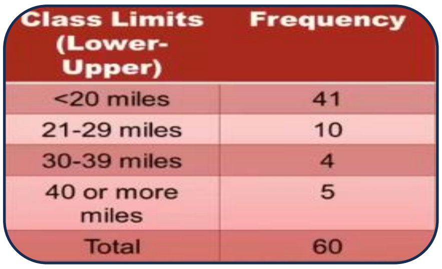
1. Classes should be the same width
2. Class width can be determined empirically
   - Example: Age 18-24, 25-34, 35-44, 45-54, 55-64, 65 and older
   - Should be based on the scientific literature
3. Can also be determined using a formula

Class Width Formula

Formula	Example
Calculate this number: maximum – minimum.	From the miles, 47 - 1 = 46.
Divide this by the number of classes desired.	If we want 6 classes, 46/6 = 7.7.
Increase this to the next whole number	We increase this up to 8

Simple Frequency Table

1. A frequency table displays each class along with the frequency (number of data points) in each class.
2. Selecting arbitrary class limits can make the frequency table unbalanced.
3. But not following the scientific literature can make your results non-comparable



Class Limits (Lower-Upper)	Frequency
<20 miles	41
21-29 miles	10
30-39 miles	4
40 or more miles	5
Total	60

Be Careful!

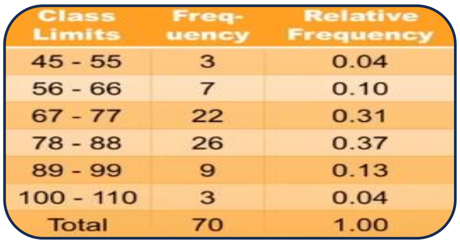
1. Make sure that all the data points are accounted for only once in one of the classes.
2. Make sure the classes cover all the data.
3. Make sure the total of your classes adds up to the total data points!

Relative Frequency Table

- "Relative" = in relationship to the rest of the data.
- Frequency = f
- Total sample size = n
- Relative frequency = f/n
- Relative frequency is the proportion of the values that are in that class.

Relative Frequency Table

- Relative frequency is something very useful to put in a frequency table.
- See how easy it is to calculate - take each class frequency divided by total.



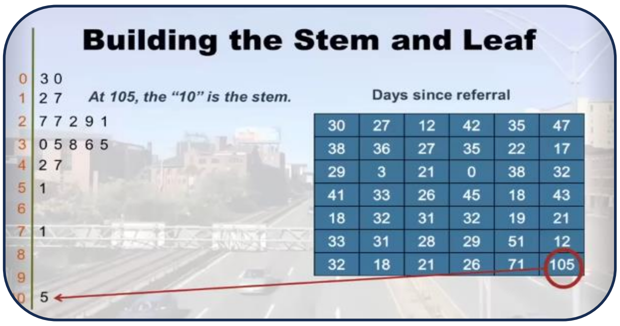
Class Limits	Freq-uency	Relative Frequency
45 - 55	3	0.04
56 - 66	7	0.10
67 - 77	22	0.31
78 - 88	26	0.37
89 - 99	9	0.13
100 - 110	3	0.04
Total	70	1.00

Frequency Tables

- Frequency tables are necessary for organizing quantitative data.
- Class width must be selected, and lower- and upper-class limits determined
- Frequencies are then filled in.
- You can also include relative frequencies.

Why is it Called Stem and Leaf?

- In a stem and leaf, there is always a "stem"
- Leaves are then added to the stem as we tally up the length of the leaves.
- Making one will help you understand the terminology.



Building the Stem and Leaf

```
0 | 3 0
1 | 2 7
2 | 7 7 2 9 1
3 | 0 5 8 6 5
4 | 2 7
5 | 1
6 |
7 | 1
8 |
9 |
0 | 5
```

At 105, the "10" is the stem.

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Organizing Quantitative Data

Frequency Table	Stem and Leaf
1. Need to set up classes, class widths	1. Do not need to set up classes or class widths
2. Need to count frequencies in each class	2. 2. No need to count. Can tally the data as you go through the list.
3. Lots of pre-calculations	3. Quicker to do

Stem-and-Leaf

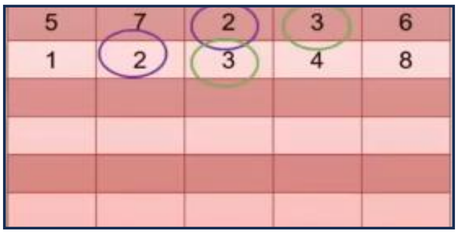
1. A stem and leaf is another way to organize quantitative data.
2. A stem and leaf is easier to make than a frequency table and requires less preparation
3. Can help you put data in order to create a frequency table

Measures of Central Tendency

1. Mode
2. Median
   • With odd number of values
   • With even number of values
3. Mean
   • Trimmed mean
   • Weighted average

Mode

1. The value that occurs most frequently in the dataset is the mode
2. It is possible to have no mode.
3. It is possible to have more than one mode.
4. Can get confusing with a lot of numbers in a short range - which is most numerous?
5. Less confusing when scale is large
   • Systolic blood pressure
   • Platelet count



5	7	2	3	6
1	2	3	4	8

What Does the Mode Tell You?

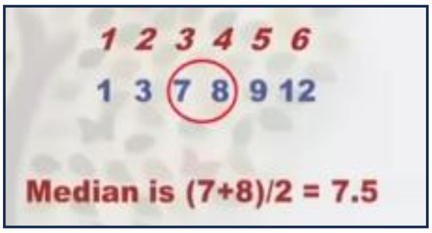
• Not much
• The most "popular" answer
• The most common result
• Not used a lot in healthcare

Median is the Center of the Data

1. Every set of quantitative data can be sorted in order of lowest to highest.
2. Sometimes there are repeats in the data
3. Sometimes there are outliers
4. Sometimes all the data values are almost the same
5. Even so, they can be arranged in order
6. The "median" is the number that is halfway

How to find the median

1. Order the data from smallest to largest
2. If there is an odd number of values, the median is the middle data value.
3. If there is an even number of values, the median is the sum of the middle two values divided by 2



1 2 3 4 5 6

1 3 7 8 9 12

Median is (7+8)/2 = 7.5

How to Find the Middle Number

1. After arranging the values in order, you have an "ordered set".
2. If there are an odd number of values (n), take n + 1, and divide by 2. Count up that many, and that is the median.

Example

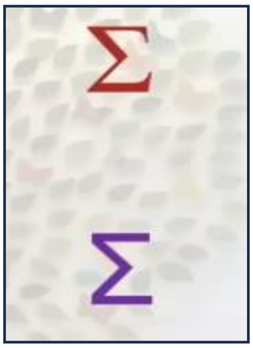
Imagine n = 21 21+1= 22, and 22/2 = 11

You would count from the beginning of the ordered set, and the 11th value in it would be the median.

What Does the Median Tell You?

1. The 50th percentile of the data.
2. The middle rank of the data.
3. The median doesn't care much about the ends of the data. Outliers don't bother it.
4. It is resistant. It is stable.

Greek Letter Capital Sigma

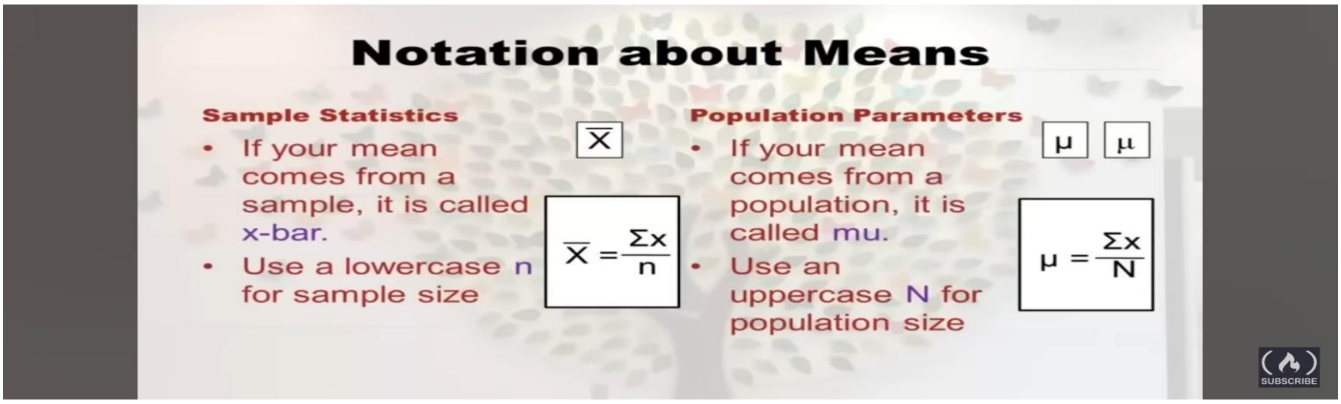


1. Whenever you see this, say in your head, "Sum of ___
2. Σx is pronounced, "Sum of x"
3. x's are the values in your data
4. Means "add up all the x's"
5. Another example:
   - Σχγ.
   - "Sum of xy."
   - Means you must have a bunch of xy's and need to add them up.
6. ∑ used a lot in statistics

Formula for Mean

1. ∑x = add up all the x's.
2. n = number of values in your data
3. After summing up all the x's, you divide by n.

Σχ/η



Notation about Means

Sample Statistics

- If your mean comes from a sample, it is called x-bar.
- Use a lowercase n for sample size

$\bar{X}$

$$\bar{X} = \frac{\Sigma x}{n}$$

Population Parameters

- If your mean comes from a population, it is called mu.
- Use an uppercase N for population size

μ μ

$$\mu = \frac{\Sigma x}{N}$$

Means VS. Medians

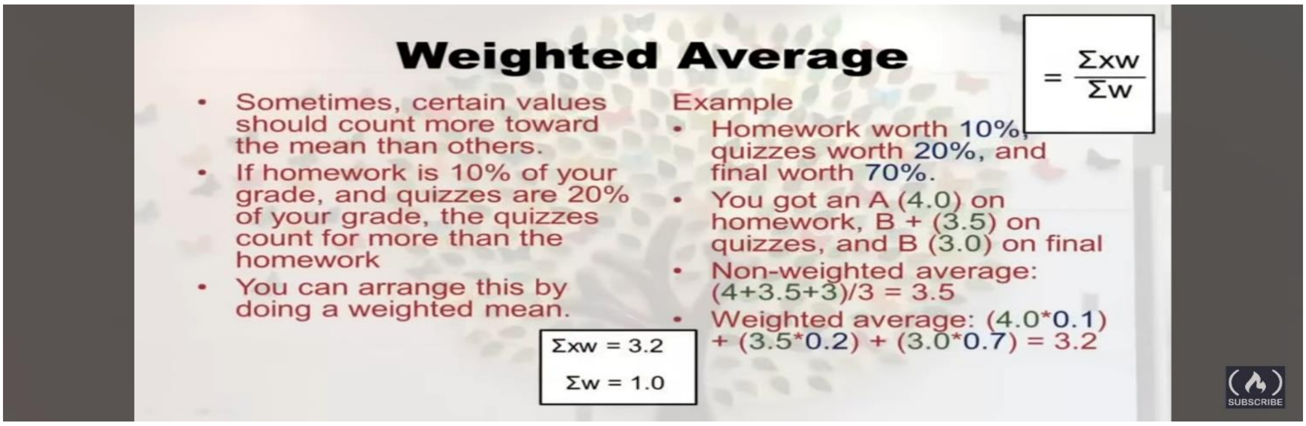
Medians	Means
Very resistant to outliers.	Not resistant to outliers.
Very stable	Not very stable.

Trimming the Mean

1. A very high value or very low value (outliers) can really throw off the mean.
2. This is not a problem with the median.
3. One solution to make the mean more resistant is to "trim" data off each end so the outliers get cut off.
4. If you trim, you have to be fair and trim the same amount off each side.

- How to make a 5% Trimmed Mean

1. Figure out how many data points you have. Then, figure out what 5% of them would be.
   - If you have 100 data points, 5% would be 5 data points.
2. Put the data in order.
3. Remove 5% from the top and 5% from the bottom.
   - Remove the 5 top ones and the 5 bottom ones.
4. Now make the mean out of the remaining data.



Weighted Average

$$= \frac{\Sigma xw}{\Sigma w}$$

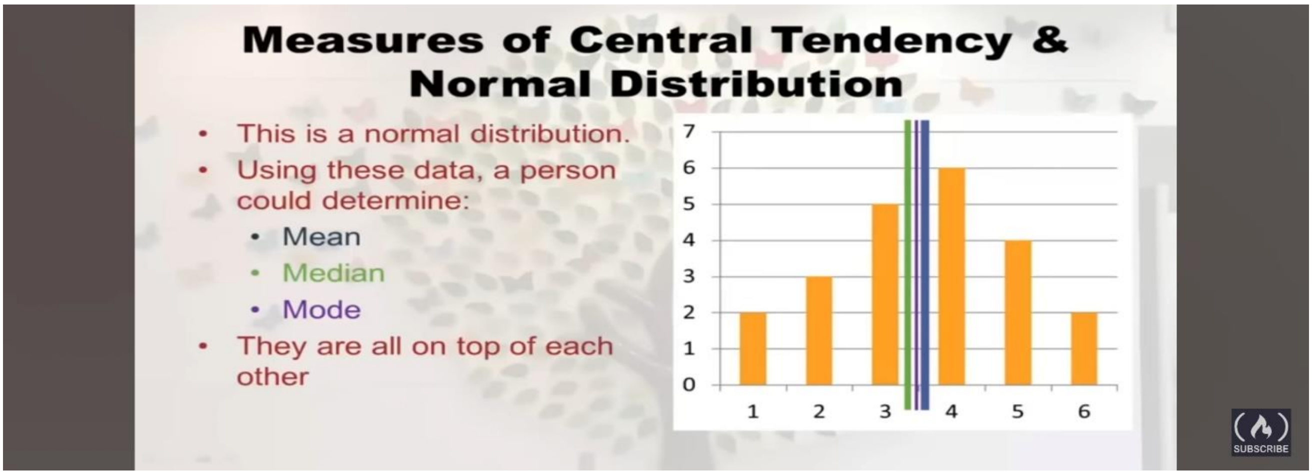
- Sometimes, certain values should count more toward the mean than others.
- If homework is 10% of your grade, and quizzes are 20% of your grade, the quizzes count for more than the homework
- You can arrange this by doing a weighted mean.

$\Sigma xw = 3.2$

$\Sigma w = 1.0$

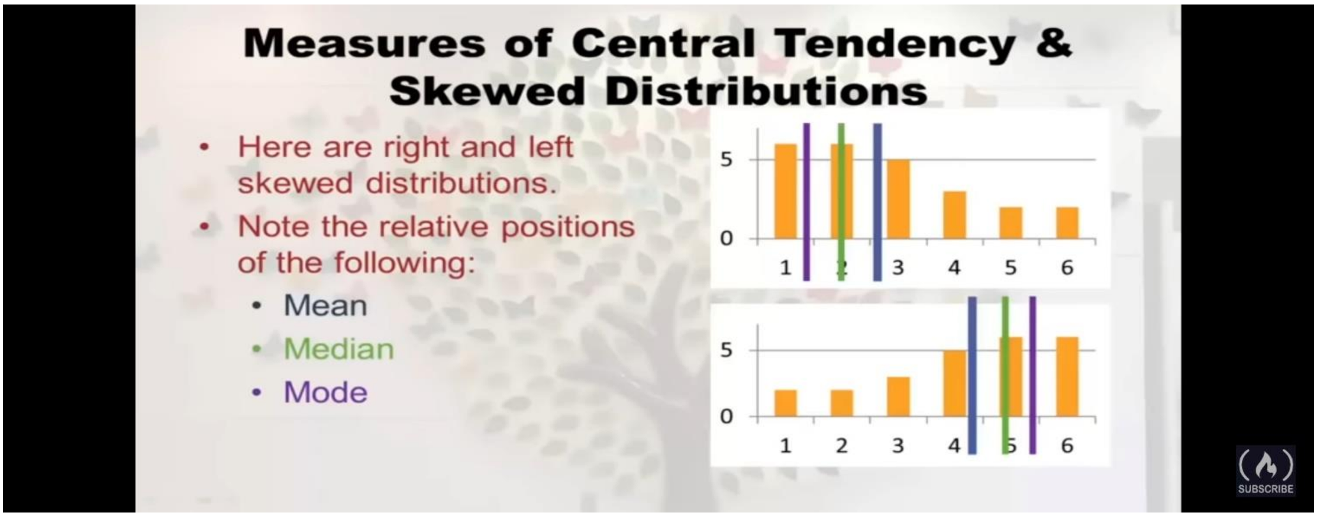
Example

- Homework worth 10%, quizzes worth 20%, and final worth 70%.
- You got an A (4.0) on homework, B + (3.5) on quizzes, and B (3.0) on final
- Non-weighted average: (4+3.5+3)/3 = 3.5
- Weighted average: (4.0*0.1) + (3.5*0.2) + (3.0*0.7) = 3.2



Measures of Central Tendency & Normal Distribution

- This is a normal distribution.
- Using these data, a person could determine:
  - Mean
  - Median
  - Mode
- They are all on top of each other



Measures of Central Tendency & Skewed Distributions

- Here are right and left skewed distributions.
- Note the relative positions of the following:
  - Mean
  - Median
  - Mode

Measures of Variation

1. Range
2. Variance and Standard Deviation
   - Sample and population
3. Coefficient of variation

Range

1. The range is the difference between the maximum and minimum value.
2. It is easy to calculate
3. Don't forget to actually do the subtraction

42 33 21 78 62

- In the above data, 78 is the maximum.

- 21 is the minimum.

- 78 minus 21 = 57

- 57 is the range

The Range is Not Very Useful

1. The range is not very useful for looking at variation because it just relies on two points.
2. It's not stable or resistant.
3. Therefore, other measures are used to look at variation typically.

    ↓ ↓

    42 33 21 78 62

    - We had a range of 78 minus 21 = 57

    - We could just change these two numbers and get a totally different range

Variance & Standard Deviation are Friends

- They are friends because this is how you calculate them:

    1. First, calculate the variance.

    2. Then, take the square root of variance, and that is the standard deviation.

- "Variance"

    - How much the data vary.

    - Think: how well does the mean represent the spread of the data?

- "Standard deviation"

    - "Standard" - following a standard, same

    - "deviation" - like a deviated septum

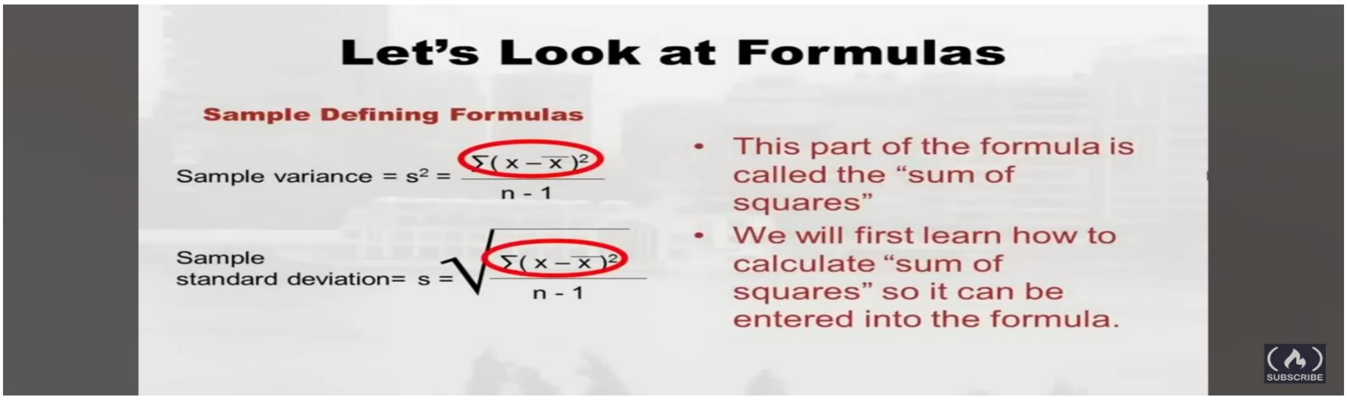
Formulas for Variance & Standard Deviation

- The formulas for sample variance and sample standard deviation are different than those for population variance and population standard deviation

    - We don't use the population ones that often

    - We will concentrate on how to use the sample ones

- Two different ways of doing the formula (for both sample and for population) - the "defining formula" and the "computational formula"
  - Both get the same results
  - I dislike the "computational formula" because I get really confused
  - Therefore, I will teach you the "defining formula"



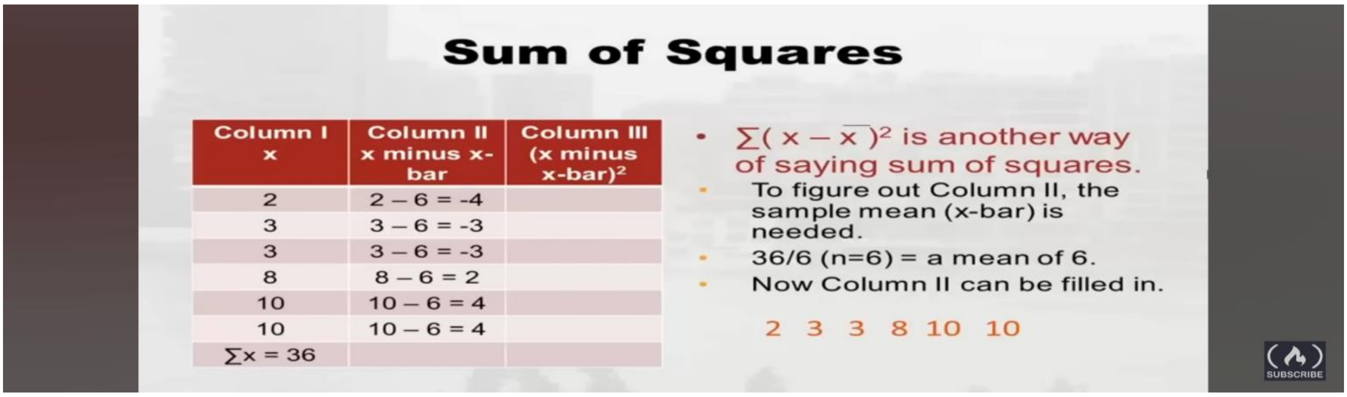
Let's Look at Formulas

Sample Defining Formulas

Sample variance = $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$

Sample standard deviation= $s = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}}$

- This part of the formula is called the "sum of squares"
- We will first learn how to calculate "sum of squares" so it can be entered into the formula.

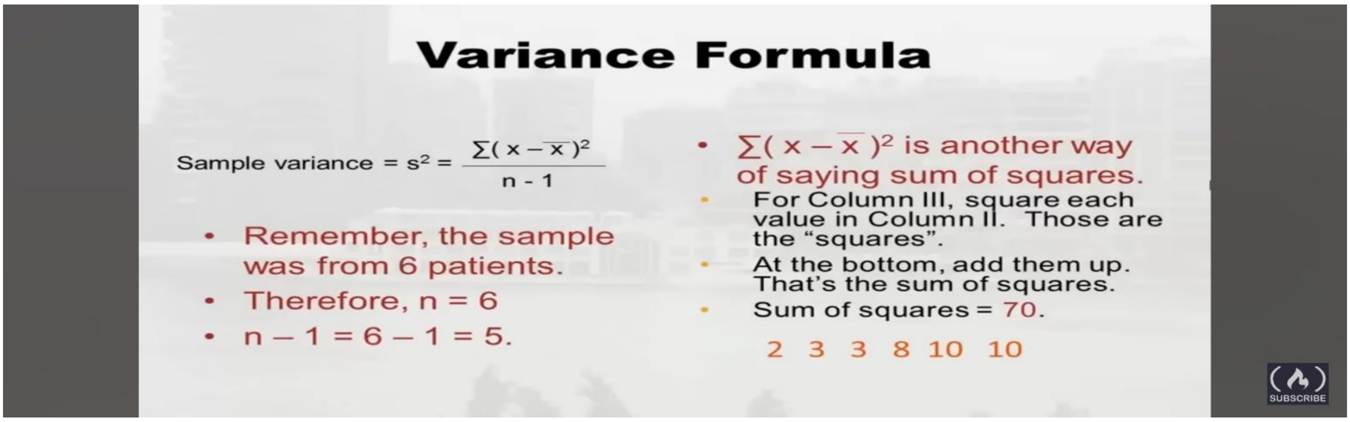


Sum of Squares

Column I x	Column II x minus x-bar	Column III (x minus x-bar)^2
2	2 – 6 = -4	
3	3 – 6 = -3	
3	3 – 6 = -3	
8	8 – 6 = 2	
10	10 – 6 = 4	
10	10 – 6 = 4	
$\sum$x = 36		

- $\sum(x-\bar{x})^2$ is another way of saying sum of squares.
- To figure out Column II, the sample mean (x-bar) is needed.
- 36/6 (n=6) = a mean of 6.
- Now Column II can be filled in.

2 3 3 8 10 10



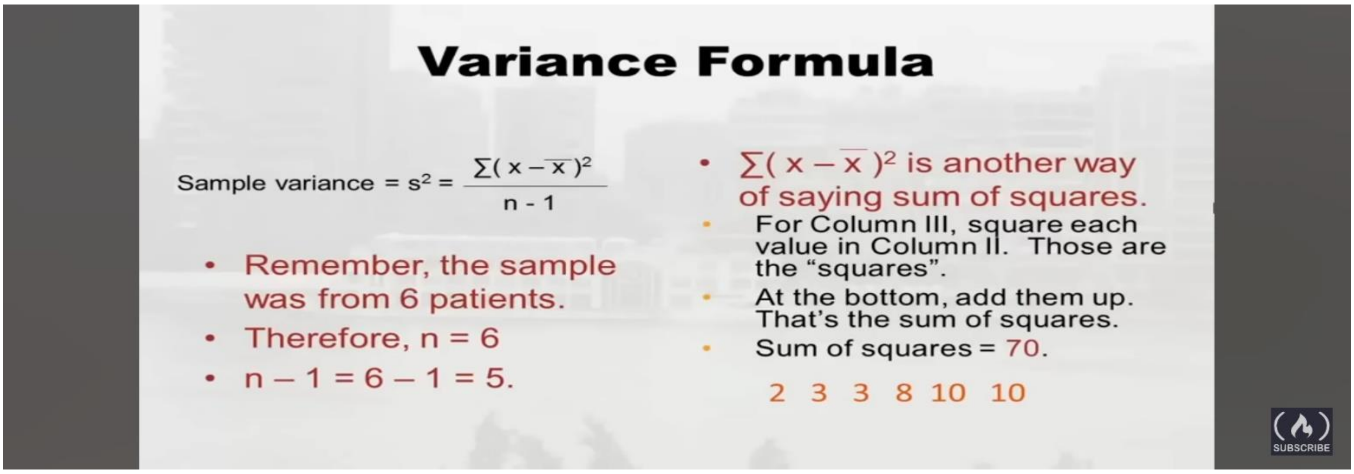
Variance Formula

Sample variance = $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$

- Remember, the sample was from 6 patients.
- Therefore, n = 6
- n – 1 = 6 – 1 = 5.

- $\sum(x-\bar{x})^2$ is another way of saying sum of squares.
- For Column III, square each value in Column II. Those are the "squares".
- At the bottom, add them up. That's the sum of squares.
- Sum of squares = 70.

2 3 3 8 10 10



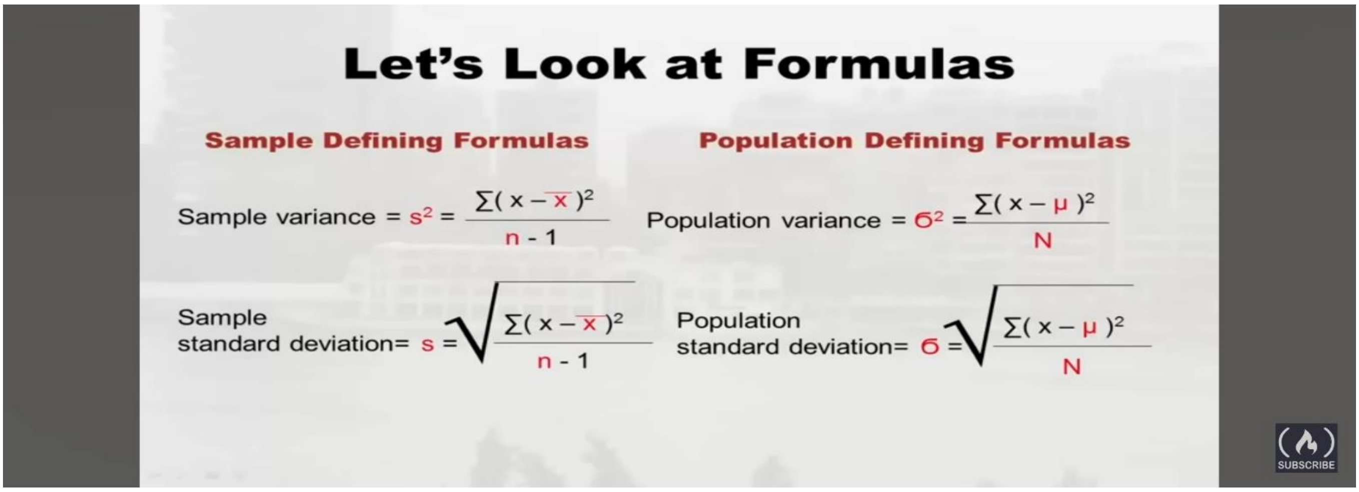
Variance Formula

Sample variance = $s^2 = \dfrac{\sum(x-\bar{x})^2}{n-1}$

- Remember, the sample was from 6 patients.
- Therefore, n = 6
- n – 1 = 6 – 1 = 5.

- $\sum(x-\bar{x})^2$ is another way of saying sum of squares.
- For Column III, square each value in Column II. Those are the "squares".
- At the bottom, add them up. That's the sum of squares.
- Sum of squares = 70.

2 3 3 8 10 10



Let's Look at Formulas

Sample Defining Formulas

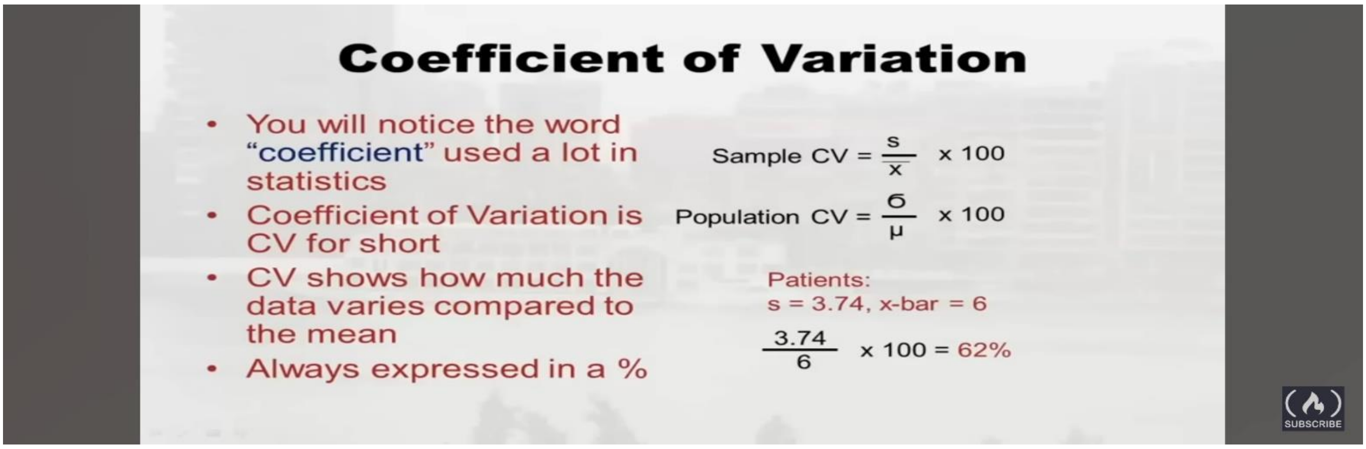
Sample variance = $s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$

Sample standard deviation= $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

Population Defining Formulas

Population variance = $\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$

Population standard deviation= $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$



Coefficient of Variation

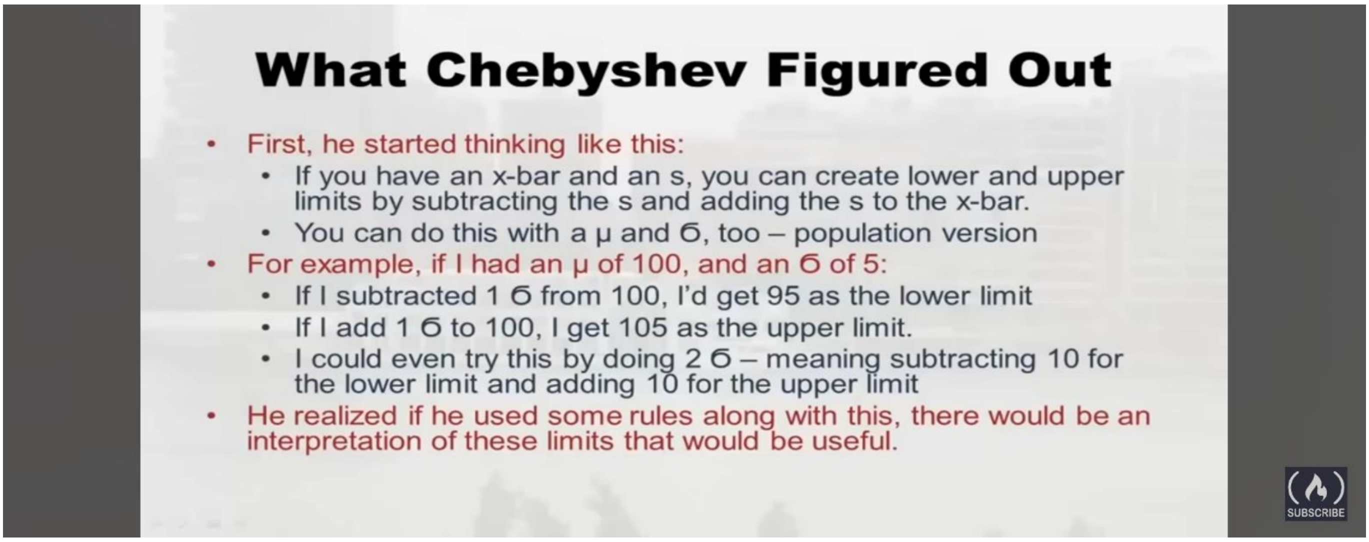
- You will notice the word "coefficient" used a lot in statistics
- Coefficient of Variation is CV for short
- CV shows how much the data varies compared to the mean
- Always expressed in a %

Sample CV = $\frac{s}{\bar{x}} \times 100$

Population CV = $\frac{\sigma}{\mu} \times 100$

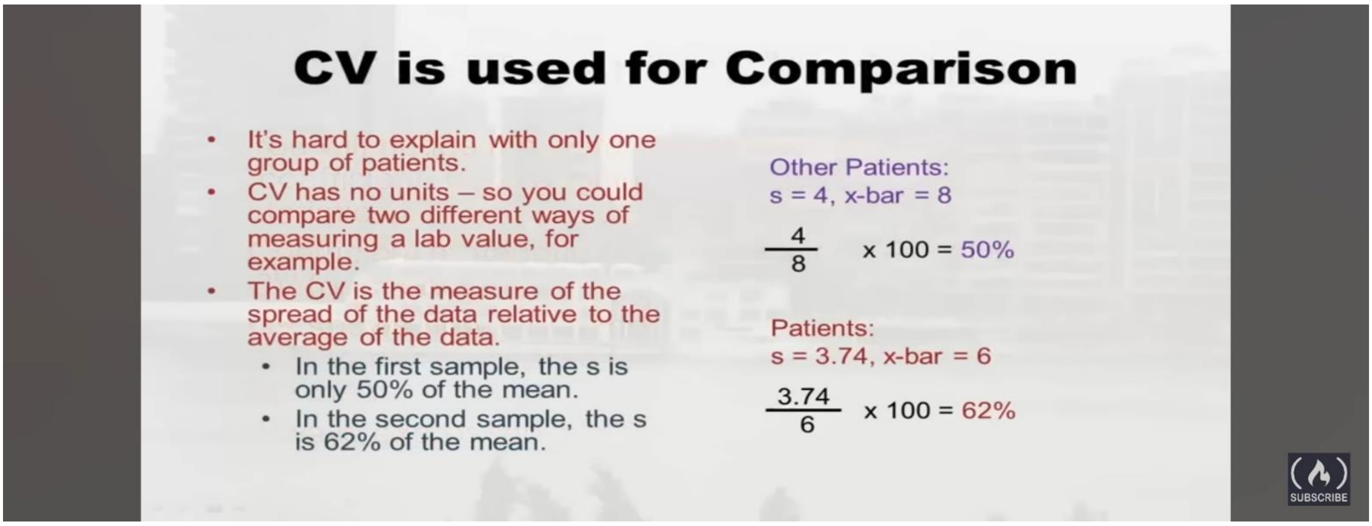
Patients:
s = 3.74, x-bar = 6

$\frac{3.74}{6} \times 100 = 62\%$



What Chebyshev Figured Out

- First, he started thinking like this:
  - If you have an x-bar and an s, you can create lower and upper limits by subtracting the s and adding the s to the x-bar.
  - You can do this with a μ and σ, too – population version
- For example, if I had an μ of 100, and an σ of 5:
  - If I subtracted 1 σ from 100, I'd get 95 as the lower limit
  - If I add 1 σ to 100, I get 105 as the upper limit.
  - I could even try this by doing 2 σ – meaning subtracting 10 for the lower limit and adding 10 for the upper limit
- He realized if he used some rules along with this, there would be an interpretation of these limits that would be useful.



CV is used for Comparison

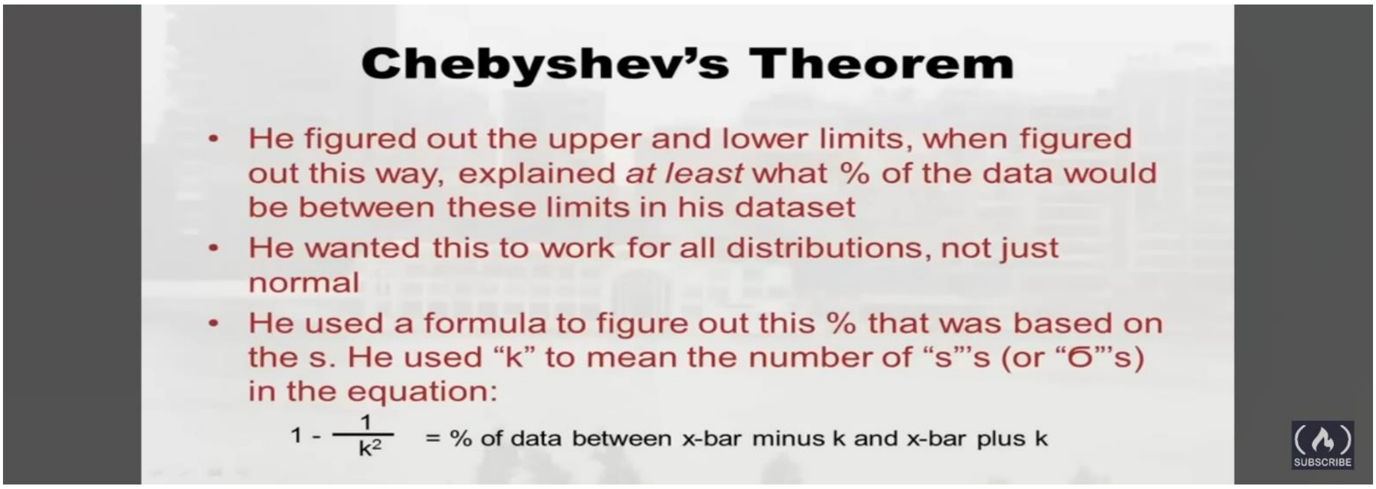
- It's hard to explain with only one group of patients.
- CV has no units – so you could compare two different ways of measuring a lab value, for example.
- The CV is the measure of the spread of the data relative to the average of the data.
  - In the first sample, the s is only 50% of the mean.
  - In the second sample, the s is 62% of the mean.

Other Patients:
s = 4, x-bar = 8

$\frac{4}{8} \times 100 = 50\%$

Patients:
s = 3.74, x-bar = 6

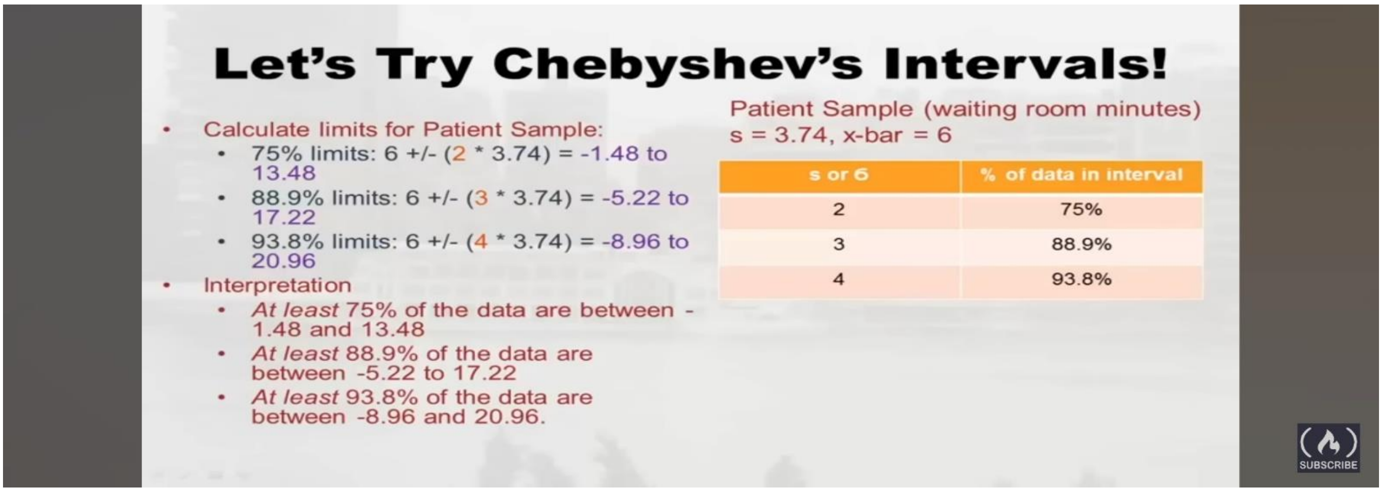
$\frac{3.74}{6} \times 100 = 62\%$



Chebyshev's Theorem

- He figured out the upper and lower limits, when figured out this way, explained *at least* what % of the data would be between these limits in his dataset
- He wanted this to work for all distributions, not just normal
- He used a formula to figure out this % that was based on the s. He used “k” to mean the number of “s”’s (or “σ”’s) in the equation:

$$1 - \frac{1}{k^2} \quad = \text{\% of data between x-bar minus k and x-bar plus k}$$

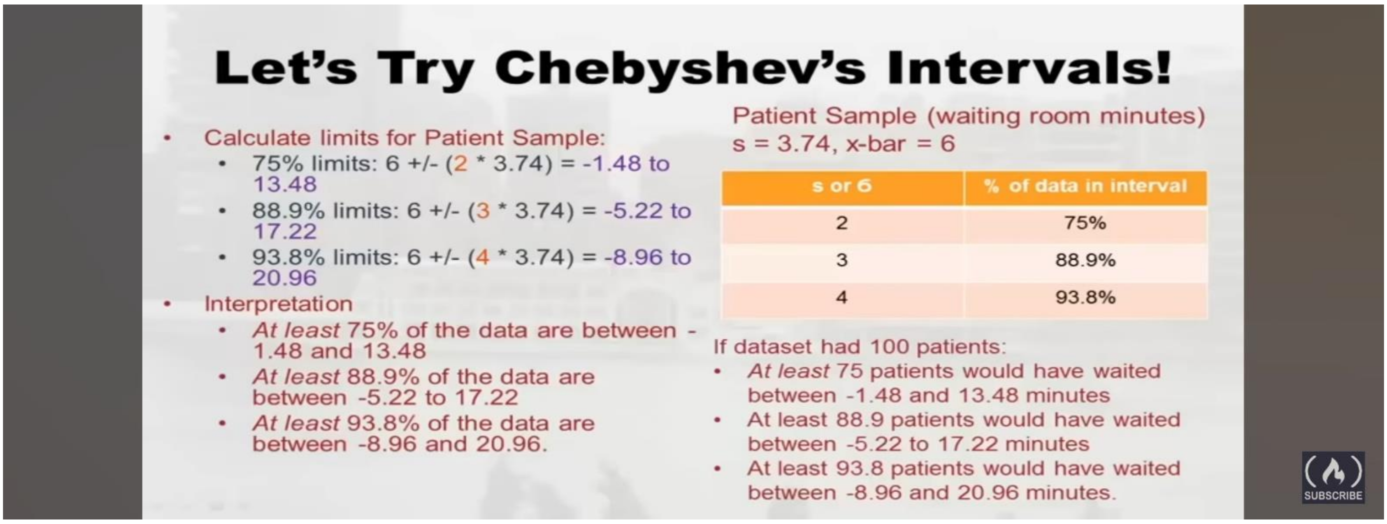


Let's Try Chebyshev's Intervals!

- Calculate limits for Patient Sample:
  - 75% limits: 6 +/- (2 * 3.74) = -1.48 to 13.48
  - 88.9% limits: 6 +/- (3 * 3.74) = -5.22 to 17.22
  - 93.8% limits: 6 +/- (4 * 3.74) = -8.96 to 20.96
- Interpretation
  - *At least* 75% of the data are between -1.48 and 13.48
  - *At least* 88.9% of the data are between -5.22 to 17.22
  - *At least* 93.8% of the data are between -8.96 and 20.96.

Patient Sample (waiting room minutes)
s = 3.74, x-bar = 6

s or σ	% of data in interval
2	75%
3	88.9%
4	93.8%



Let's Try Chebyshev's Intervals!

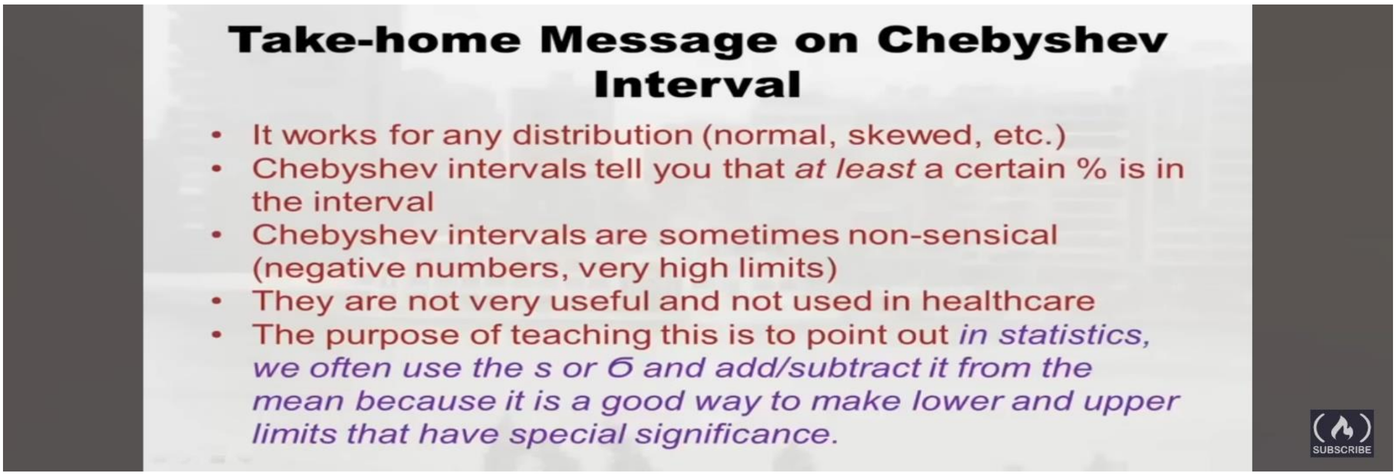
- Calculate limits for Patient Sample:
  - 75% limits: 6 +/- (2 * 3.74) = -1.48 to 13.48
  - 88.9% limits: 6 +/- (3 * 3.74) = -5.22 to 17.22
  - 93.8% limits: 6 +/- (4 * 3.74) = -8.96 to 20.96
- Interpretation
  - *At least* 75% of the data are between -1.48 and 13.48
  - *At least* 88.9% of the data are between -5.22 to 17.22
  - *At least* 93.8% of the data are between -8.96 and 20.96.

Patient Sample (waiting room minutes)
s = 3.74, x-bar = 6

s or σ	% of data in interval
2	75%
3	88.9%
4	93.8%

If dataset had 100 patients:

- *At least* 75 patients would have waited between -1.48 and 13.48 minutes
- At least 88.9 patients would have waited between -5.22 to 17.22 minutes
- At least 93.8 patients would have waited between -8.96 and 20.96 minutes.



Take-home Message on Chebyshev Interval

- It works for any distribution (normal, skewed, etc.)
- Chebyshev intervals tell you that *at least* a certain % is in the interval
- Chebyshev intervals are sometimes non-sensical (negative numbers, very high limits)
- They are not very useful and not used in healthcare
- The purpose of teaching this is to point out *in statistics, we often use the s or σ and add/subtract it from the mean because it is a good way to make lower and upper limits that have special significance.*

What are Percentiles?

- Quantitative data
- Remember standardized tests...
- Example: If you test at the 77th percentile, it means you did better than 77% of the people taking the test.
- If 100 people took the test, you'd have done better than 77 of them.

Percentile Definition

1. Percentiles can be between 1 and 99
- You can't have a -2nd percentile, or a 105th percentile
2. Whatever number you pick:
- That % of values fall below the number
- And 100 minus that % of values fall above the number
3. Example: 20 people take a test.
- Let's say there is a maximum score of 5 on the test.
- The 25th percentile means 25% of the scores fall below this score, and 75% fall above that score.
4. Let's say it is an easy test, and 12 people get a 4, and the remaining 8 get a 5. The 25th percentile, or the score the cuts off the bottom 5 tests scores, will be 4. (Even the 50th percentile will be 4.)
5. This would come out very different if it were a hard test, and most people got below a score of 3.

Quartiles

1. Quartiles is a specific set of percentiles

• 1st quartile: 25th percentile

• 2nd quartile: 50th percentile (also median!)

• 3rd quartile: 75th percentile

2. These can be calculated by hand.

Computing Quartiles

1. Order the data from smallest to largest.

2. Find the median.

• 2nd quartile	• 50th percentile

3. Find the median of the lower half of the data.

• 1st quartile	• 25th percentile

4. Find the median of the upper half of the data.

• 3rd quartile	• 75th percentile

• Remember the range?

• New! Interquartile range

• Once you have 3rd quartile and 1st quartile you can calculate interquartile range (IQR)

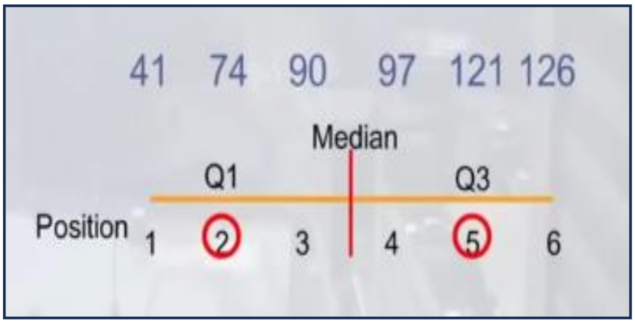
• 3rd quartile minus 1st quartile = IQR

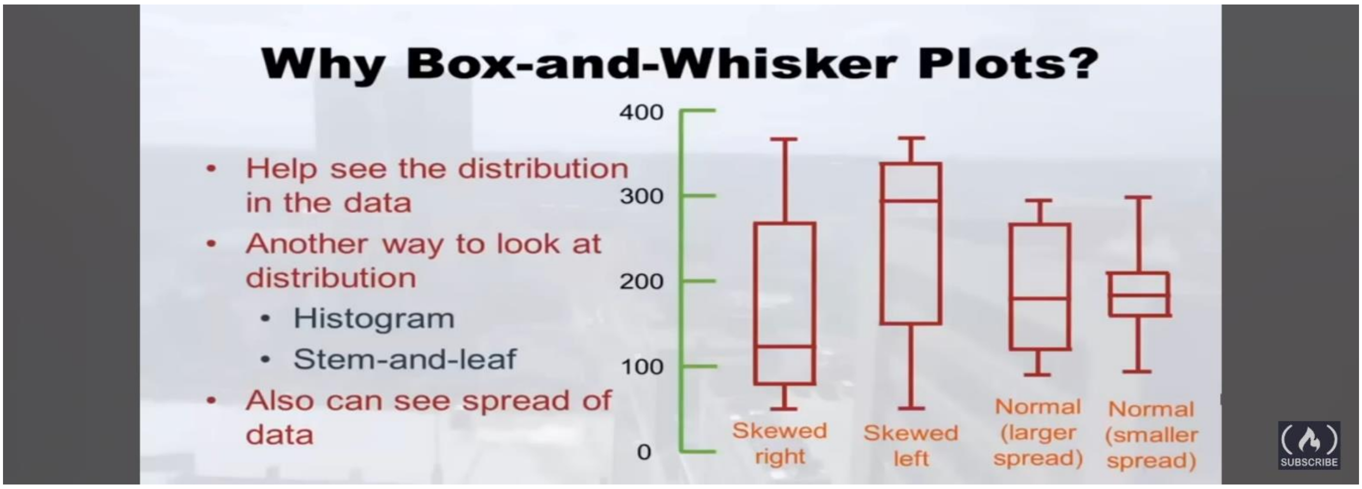
More Notes on Q1 and Q3

1. Imagine we started with 6 values.

• The median would be between the 3rd and 4th position.

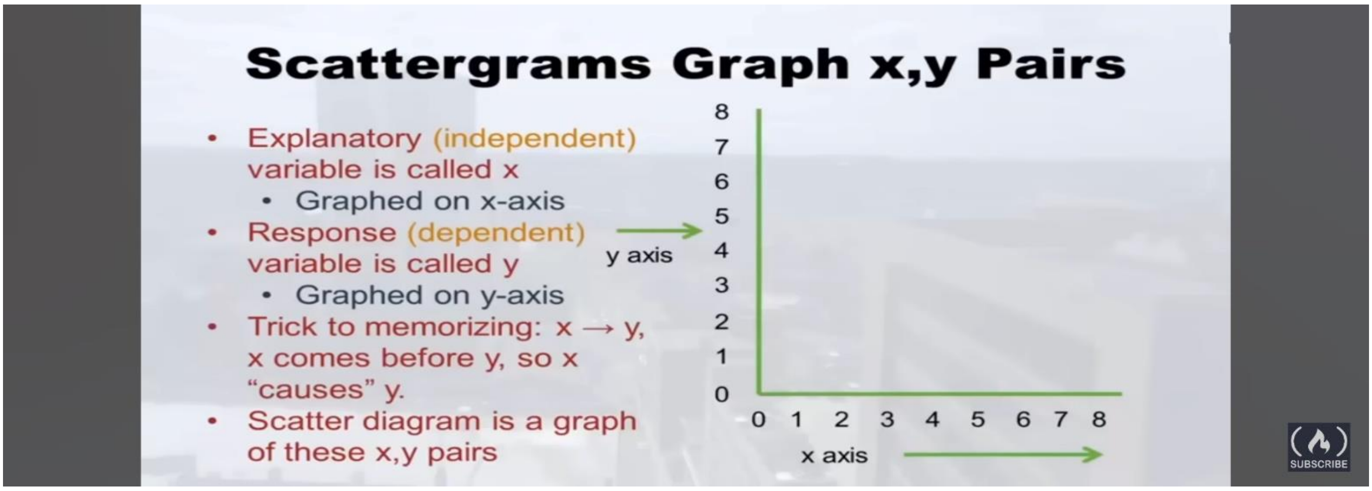
• Therefore, all 3 values below the median would be Considered in calculating Q1, and all 3 above the median would be considered in calculating Q3.





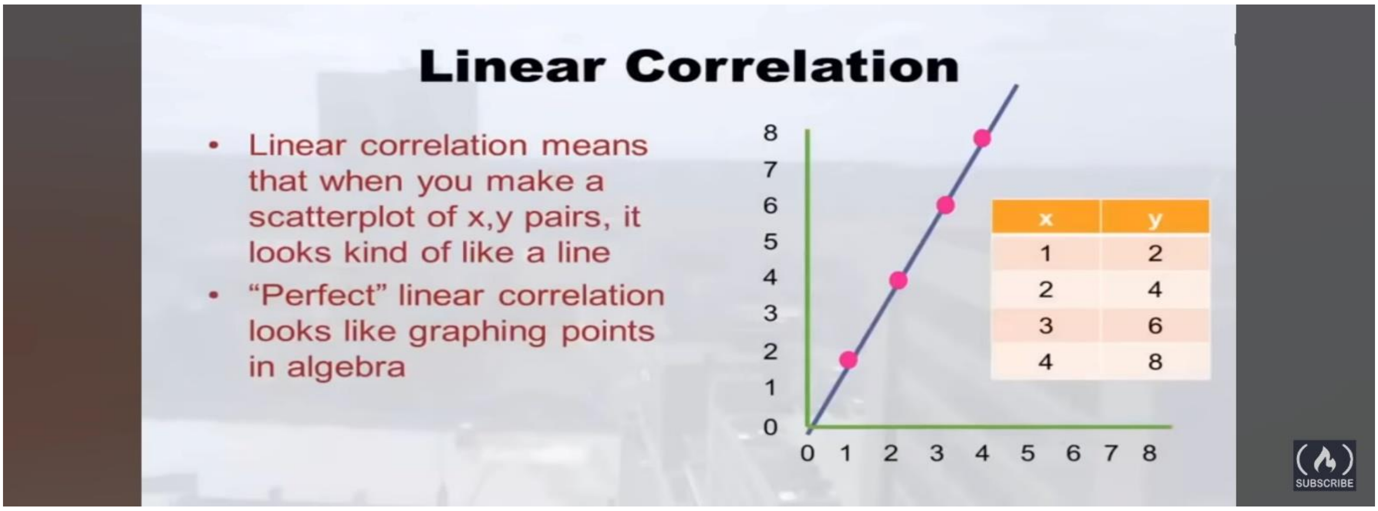
Why Box-and-Whisker Plots?

- Help see the distribution in the data
- Another way to look at distribution
  - Histogram
  - Stem-and-leaf
- Also can see spread of data



Scattergrams Graph x,y Pairs

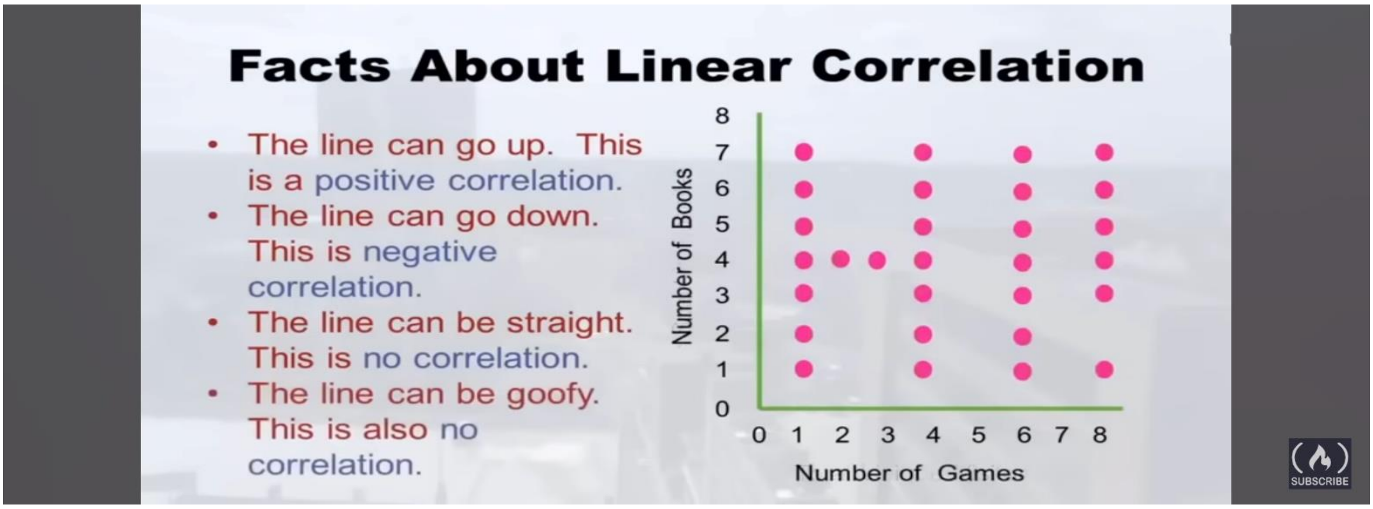
- Explanatory (independent) variable is called x
  - Graphed on x-axis
- Response (dependent) variable is called y
  - Graphed on y-axis
- Trick to memorizing: x → y, x comes before y, so x "causes" y.
- Scatter diagram is a graph of these x,y pairs



Linear Correlation

- Linear correlation means that when you make a scatterplot of x,y pairs, it looks kind of like a line
- "Perfect" linear correlation looks like graphing points in algebra

x	y
1	2
2	4
3	6
4	8



Facts About Linear Correlation

- The line can go up. This is a positive correlation.
- The line can go down. This is negative correlation.
- The line can be straight. This is no correlation.
- The line can be goofy. This is also no correlation.

Correlation Has Two Attributes

Direction

- Positive correlation
- Negative correlation
- No correlation

Strength

- Strength refers to how close to the line all the dots fall.
- If they fall really close to the line, it is strong
- If they fall kind of close to the line, it is moderate
- If they aren't very close to the line, it is weak

Outliers in Correlation

- Outliers can have a very powerful effect on a correlation
- An outlier in any of the 4 corners of the plot can really affect the direction of the line
- An outlier can also change the correlation from strong and moderate to weak
- It's good to look at a scatterplot to make sure you identify outliers

Correlation Coefficient r

- Remember "coefficient" from CV (coefficient of variation)?
- Coefficient just means a number
- r stands for the sample correlation coefficient
    - Remember! Corrrrrrrrrrrrrrrrrrelation
    - Population correlation coefficient = **P**
- We will only focus on r

What is r?

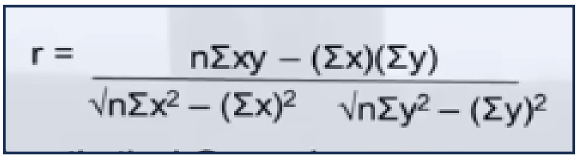
What it is

- A numerical quantification of how correlated a set of x,y pairs are
- Calculated from plugging x,y pairs into an equation
- Has a defining formula and a computational formula
- I will demonstrate computational formula

How to interpret it

- The r calculation produces a number
- The lowest number possible is -1.0
    - Perfect negative correlation
- The highest possible number is 1.0
    - Perfect positive correlation
- All others are in-between

Computational Formula



$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Hypothetical Scenario

- We have 7 patients
- They have come to the clinic for appointments throughout the year.
- We predict those with a higher diastolic blood pressure (DBP) will have more appointments
- We take DBP at last appointment as "x"
- We take number of appointments over the year as "y"

FLASHBACK! ...to Chapter 3.2

- Notice all the ∑'s
- As before, we will
    - make columns
    - make calculations
    - Then add up the columns to get these ∑'s

Facts About r

- r requires data with a "bivariate normal distribution" - we do not cover looking at this in this class, but please know this.
- r does not have units.
- Perfect linear correlation is r=-1.0 or r=1.0 (depending on direction). No linear correlation is r=0.
- Positive r means as x goes up, y goes up, and as x goes down, y goes down.

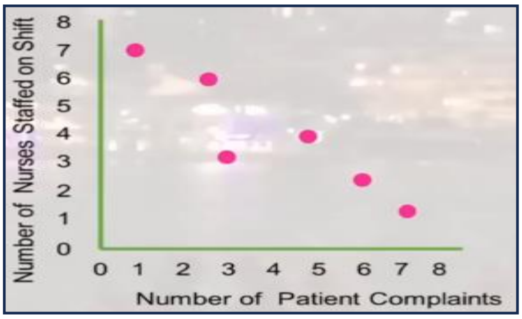
- Negative r means as x goes up, y goes down, and as x goes down, y goes up.
- Even if you switched x and y on the axes, you'd get the same r.
- Even if you converted x and y to different units (e.g., you converted measurements into the metric system), you'd get the same r.

Correlation is not Causation

- Beware of lurking variables!
- Selecting x and y is political - you are implying x could cause y
    - Example: Taller people are heavier, so x=height and y=weight
    - People who are overweight do not suddenly grow taller
- But there are other causes of weight besides height.
- Genetics can cause both height and weight.
    - A genetic profile that leads to tallness and obesity could be a lurking variable in the relationship between height and weight.

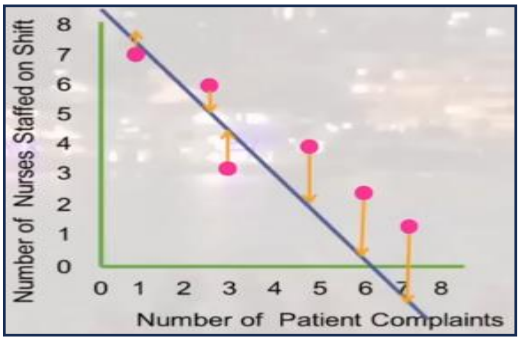
Where Does the Line Go?

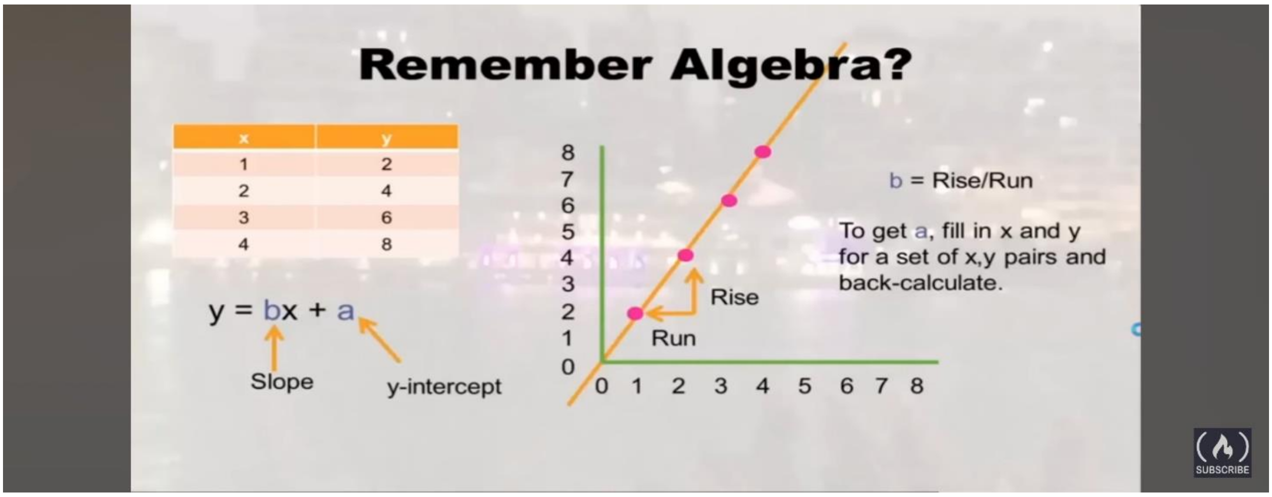
1. In the last chapter, we plotted scattergrams.
2. I just drew a line for demonstration - but there is an official rule as to where this line goes.
3. The rule is that the line has to meet the "least squares criterion"



Where Does the Line Go?

1. "Least-squares line"
    - The vertical distances between the dot and line is squared to get rid of negative sign
    - These are called "squares"
2. The line belongs where it would cause the smallest sum of squares for the whole dataset.





Remember Algebra?

x	y
1	2
2	4
3	6
4	8

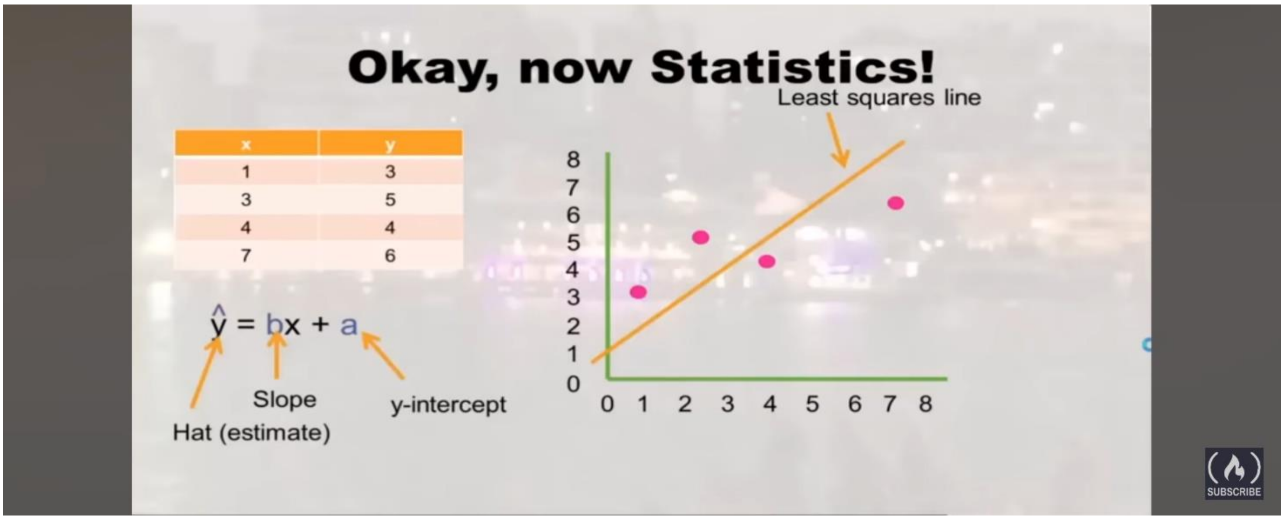
y = bx + a

Slope

y-intercept

b = Rise/Run

To get a, fill in x and y for a set of x,y pairs and back-calculate.



Okay, now Statistics!

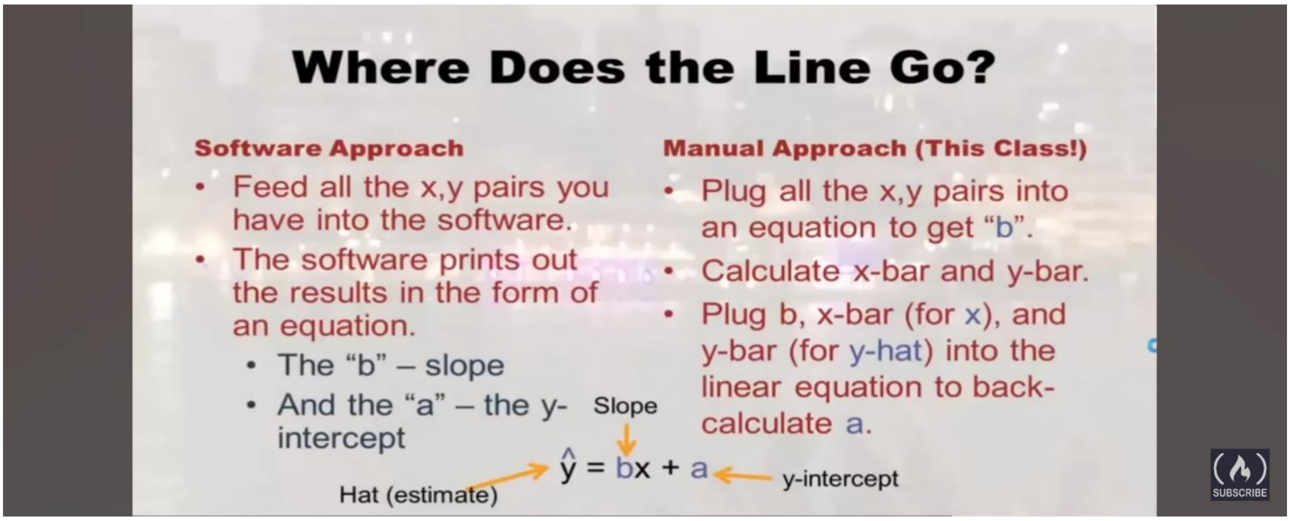
x	y
1	3
3	5
4	4
7	6

$\hat{y} = bx + a$

Hat (estimate)

Slope

y-intercept



Where Does the Line Go?

Software Approach

- Feed all the x,y pairs you have into the software.
- The software prints out the results in the form of an equation.
  - The “b” – slope
  - And the “a” – the y-intercept

Manual Approach (This Class!)

- Plug all the x,y pairs into an equation to get “b”.
- Calculate x-bar and y-bar.
- Plug b, x-bar (for x), and y-bar (for y-hat) into the linear equation to back-calculate a.

$\hat{y} = bx + a$

Hat (estimate)

Slope

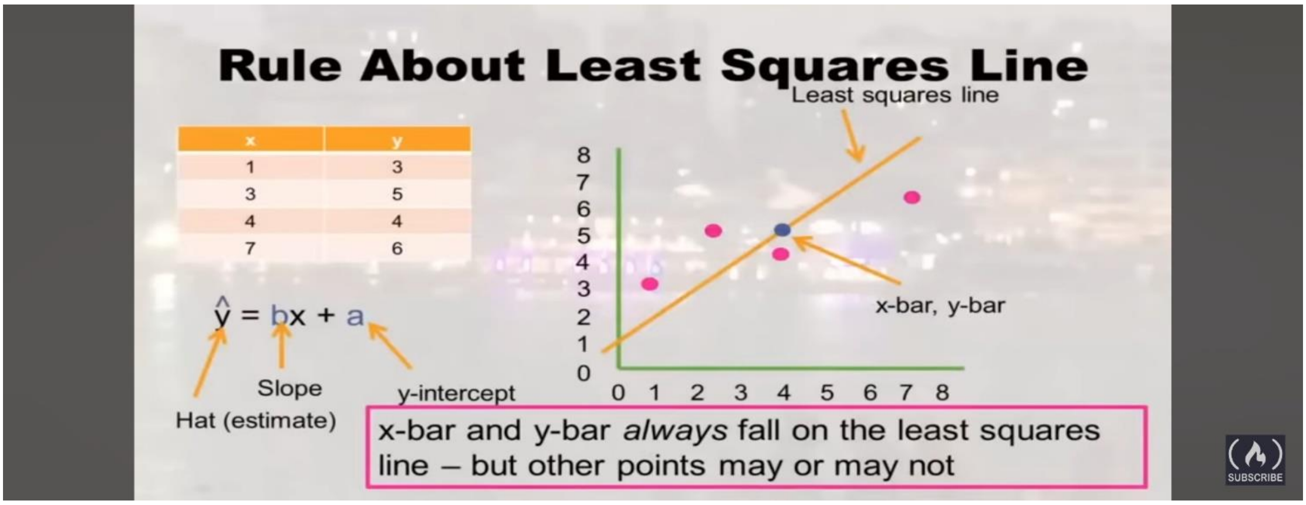
y-intercept

Recycling!

1. Least-squares line is usually done along with r
2. SAVE YOUR CALCULATIONS from r to recycle when calculating b:

   • Σχ, Σγ, $Σχ^2$, and Σxy

3. Also save your r! You will need it later for the Coefficient of Determination.
4. NOTE: You will need to calculate x-bar and y-bar - this was not done in r



Rule About Least Squares Line

x	y
1	3
3	5
4	4
7	6

$\hat{y} = bx + a$

Hat (estimate)

Slope

y-intercept

x-bar and y-bar *always* fall on the least squares line – but other points may or may not

Facts About the Slope (b)

1. The slope (b) of the least-squares line tells us how many units the response variable (y) is expected to change for each 1 unit of change in the explanatory variable (x).
2. For our example: **ŷ = 1.1x - 80.0**

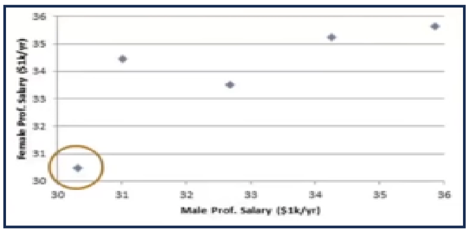
   - x=DBP, y=# of Appointments

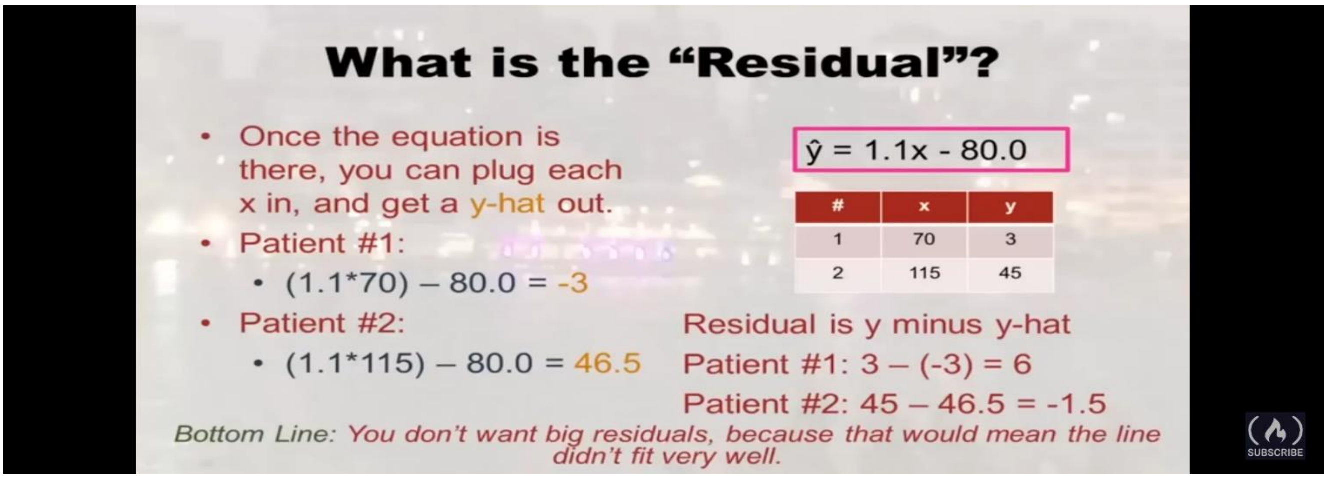
   - For each increase in 1 mmHg of DBP (x), there is a 1.1 increase in the number of appointments the patient had over the past year (y)

3. The number of units change in the y for each unit change in x is called the "marginal change" in the y.

Influential Points

1. Like with r, if a point is an outlier, it can drastically influence the least squares line equation.
2. An extremely high x or extremely low x can do that.
3. Always check the scattergram first for outliers!





What is the "Residual"?

- Once the equation is there, you can plug each x in, and get a y-hat out.
- Patient #1:
  - (1.1*70) – 80.0 = -3
- Patient #2:
  - (1.1*115) – 80.0 = 46.5

ŷ = 1.1x - 80.0

#	x	y
1	70	3
2	115	45

Residual is y minus y-hat

Patient #1: 3 – (-3) = 6

Patient #2: 45 – 46.5 = -1.5

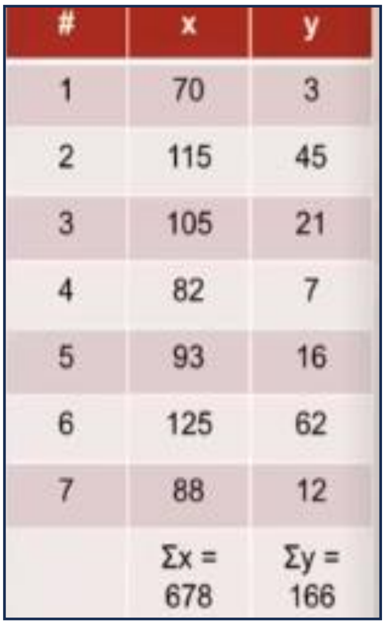
Bottom Line: You don't want big residuals, because that would mean the line didn't fit very well.

Using Least Squares Line Equation for Prediction

1. Let's say you knew someone's DBP and you wanted to predict how many appointments s/he would have next year
2. You can plug the DBP in as x, and get y-hat out, and say that's your prediction
3. If you use an x within the range of the original equation (70-125), this type of prediction is called interpolation.
4. If you use an x from outside the range (such as 65, or 130), it is extrapolation not a great idea.

Example of Interpolation

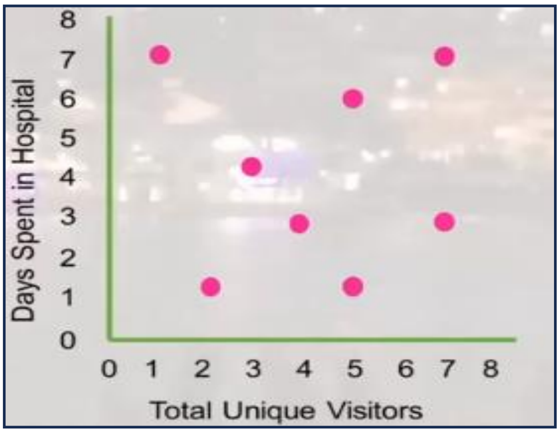
1. The patient in your study has a DBP of 80. That is within the range of your x's. Let's predict how many appointments he will have next year. Here's the equation: **ŷ = 1.1x - 80.0**
   (1.1*80)-80 = 8, so we predict this patient will come to 8 appointments next year.



#	x	y
1	70	3
2	115	45
3	105	21
4	82	7
5	93	16
6	125	62
7	88	12
	Σx = 678	Σy = 166

Is it Really This Easy to Make Predictions Using the Least Squares Line?

1. No. You can make a linear equation out of any x,y pairs.
2. If there is no linear correlation, though, the line is meaningless for prediction.
3. Imagine a line for this scatter plot - would that really work for prediction?
4. To evaluate if our least-squares line equation is should be used for interpretation, we use the Coefficient of Determination



The Coefficient of Determination (CD)

1. This is r^2 (in other words, r times r)

   - Then, like CV, we turn it into a%

2. In the example, our r=0.95
3. 0.95 0.95 = .90
4. CD = 90%
5. 90% = explained variation in y (by the linear equation)
6. 100%-90% = 10% unexplained variation
7. "90% of the variation in the number of appointments is explained by DBP."
8. "10% of the variation in the number of appointments is NOT explained by DBP."
9. What happens if the CD is low?

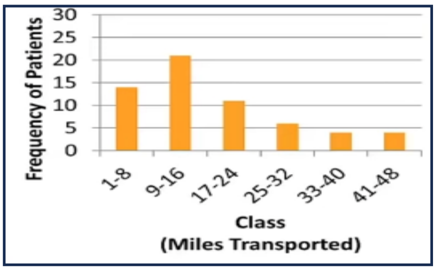
   - CD should be better than at least 50% (random)

   - The higher, the better

   - If it is low, it means other variables might be needed to explain more of the variation

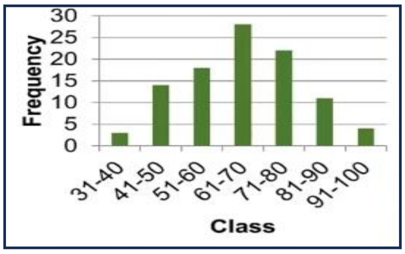
Remember Distributions?

1. Using quantitative variable
2. Classes determined
3. Frequency table made
4. Frequency histogram
5. Then we could see the distribution

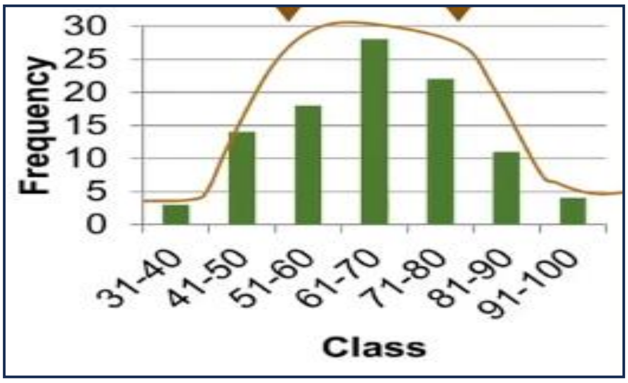


Remember the Normal Distribution?

1. Imagine a large class (n=100) takes a very difficult test
2. The test is worth 100 points, but it's so hard, no one actually gets 100 points
3. Instead, the mode is near a C grade



Properties of the Normal Curve



1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur at about mean +/- 1 sd
5. The area under the entire curve is 1 (think: 100%).

Remember Chebyshev?

1. Intervals have boundaries, or limits: lower limit and upper limit.
2. Remember Chebyshev Intervals?

   - They say, "At least % of the data fall in the interval."

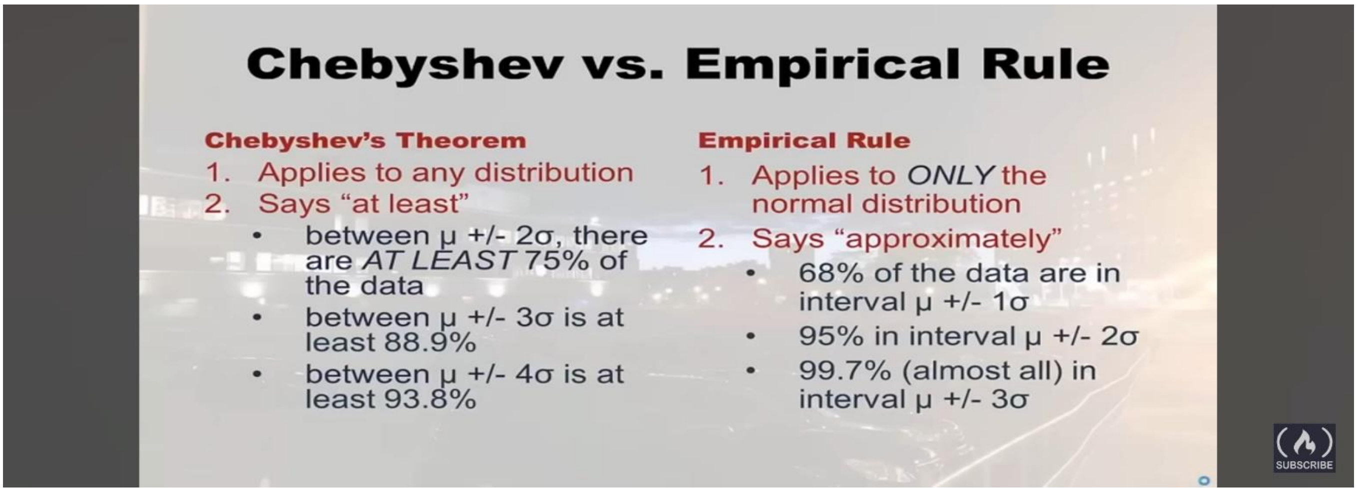
   - When lower limit was μ-20, and upper limit was μ+20, at least 75% of the data were in the interval.

3. Imagine n=100 students, μ score on test 65.5, σ = 14.5

- Lower limit: 65.5- (2*14.5) = 36.5

- Upper limit: 65.5+ (2*14.5) = 94.5

- So if you had 100 data points, at least 75 would be between 36.5 and 94.5.



Chebyshev vs. Empirical Rule

Chebyshev's Theorem

1. Applies to any distribution
2. Says "at least"
   - between μ +/- 2σ, there are *AT LEAST* 75% of the data
   - between μ +/- 3σ is at least 88.9%
   - between μ +/- 4σ is at least 93.8%

Empirical Rule

1. Applies to *ONLY* the normal distribution
2. Says "approximately"
   - 68% of the data are in interval μ +/- 1σ
   - 95% in interval μ +/- 2σ
   - 99.7% (almost all) in interval μ +/- 3σ

Conclusion

1. The Empirical Rule helps establish intervals that apply to normally distributed data
2. These intervals have a certain percentage of the data points in them
3. These intervals depend on the mean and standard deviation of the data distribution

Z-scores

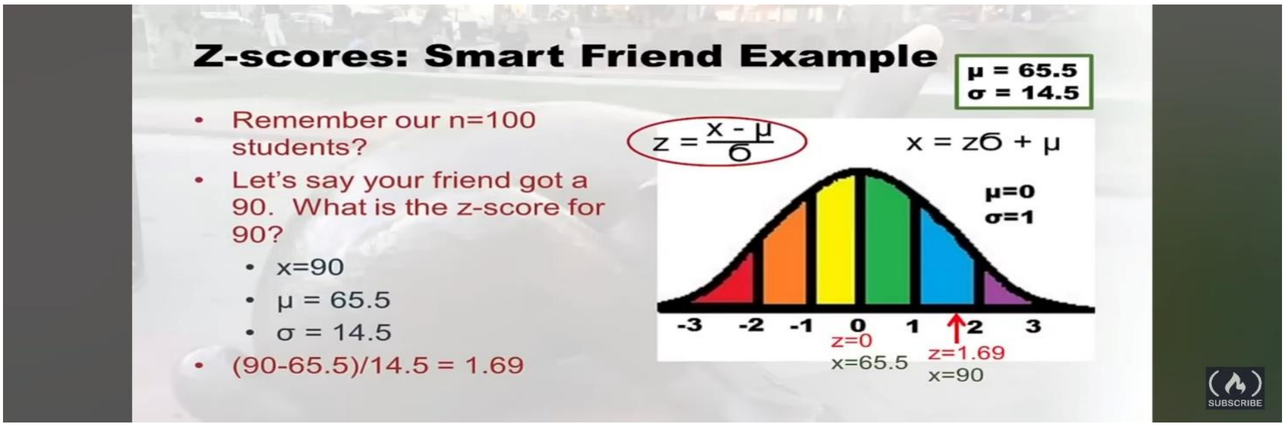
Every value on a normal distribution (every "x") can be converted to a z-score.

You must know the following to use formula:

- The "x" - what you want to convert to z

- The u of the distribution

- The o of the distribution

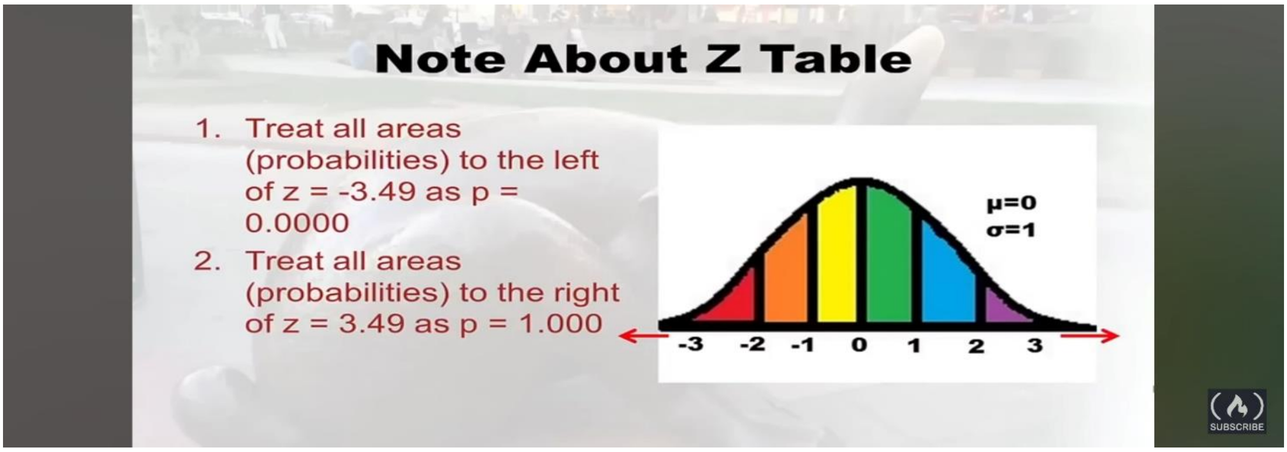


Z-scores: Smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$

- Remember our n=100 students?
- Let's say your friend got a 90. What is the z-score for 90?
  - x=90
  - μ = 65.5
  - σ = 14.5
- (90-65.5)/14.5 = 1.69

$z = \frac{x - \mu}{\sigma}$ $x = z\sigma + \mu$



Note About Z Table

1. Treat all areas (probabilities) to the left of z = -3.49 as p = 0.0000
2. Treat all areas (probabilities) to the right of z = 3.49 as p = 1.000

Z-Score Quiz

1. Where is x?

Ans Usually in the question.

2. What do you do with an x?

Ans Calculate a z-score.

3. What do you do with a z?

Ans Look it up in the Z table.

4. What if the question asks for x?

Ans Use the x formula

5. What if the question gives you a p?

Ans Dig around in the table to find the p to map back to z, then use x formula

Tips for Getting Z-Scores and Probabilities Right

1. Draw a picture: Graph out the question. Draw the curve, the line for μ, and where the x goes (above or below the μ).

   • If there is one x, shade in the part of the curve wanted (above or below).

   • If there are 2 x's, shade in the area wanted (usually in between them).

   • If it's a "calculate the x" question, put where the z or p is, and shade in the probability you are calculating.

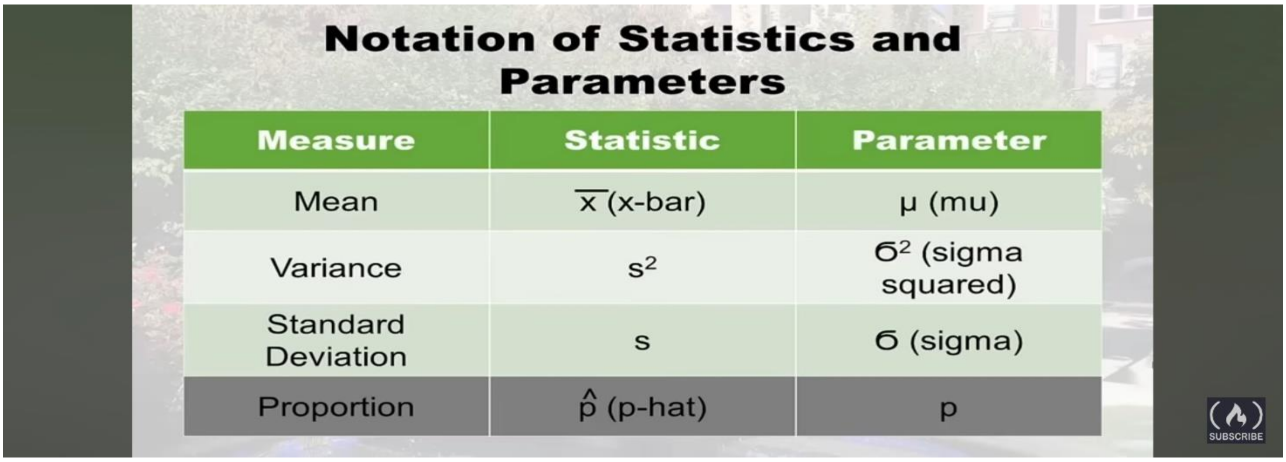
2. x is usually in the question: The question must give you μ and σ, and students usually can find those, but then they can't find the x.

3. Don't mistake little z's for p's: Sometimes a little z-score (like 0.023) looks like a p. Don't be fooled! You still have to look it up.
4. Check logic against your picture: If you shaded in a big part of your picture, your probability should be bigger than 0.5000 or 50%.

Reminder of Statistic and Parameter

• A statistic is a numerical measure describing a sample.

• A parameter is a numerical measure describing a population.



Notation of Statistics and Parameters

Measure	Statistic	Parameter
Mean	$\overline{x}$ (x-bar)	μ (mu)
Variance	s^2	σ^2 (sigma squared)
Standard Deviation	s	σ (sigma)
Proportion	$\hat{p}$ (p-hat)	p

Types of Inferences

1. Estimation: we estimate the value of a population parameter using a sample
2. Testing: we do a test to help us make a decision about a population parameter
3. Regression: we make predictions or forecasts about a statistic

Frequency vs. Sampling Distribution

Frequency Distribution

1. Make a histogram of a quantitative variable.
2. Draw the shape and name the distribution.

Sampling Distribution

1. Start with a population.
2. Decide on an n.
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

Explanation of Sampling Distribution

Sampling Distribution

1. Start with a population.
2. Decide on an n. **n=5**
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

Definition of a Sampling Distribution

1. A sampling distribution is a probability distribution of a sample statistic based on all possible simple random samples of the same size from the same population.
2. In the next section, we will talk about the Central Limit Theorem, which is a proof that shows how we can use a sampling distribution for inference.

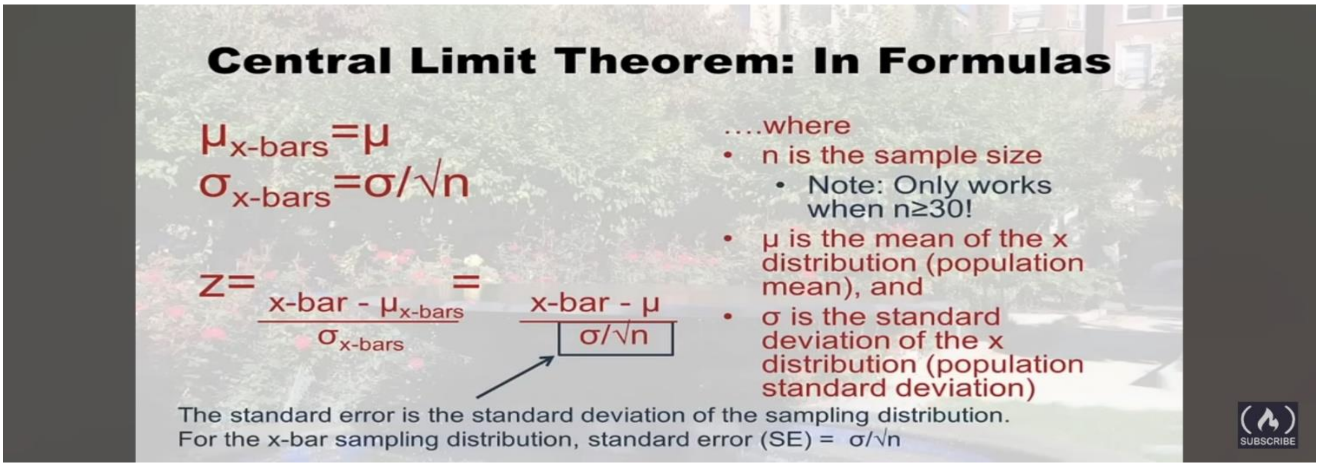
Central Limit Theorem: In Words

For any normal distribution:

1. The sampling distribution (the distributions of x-bars from all possible samples) is also a normal distribution

2. The mean of the x-bars is actually μ

3. The standard deviation of the x-bars is actually σ/vη



Central Limit Theorem: In Formulas

$\mu_{\text{x-bars}}=\mu$

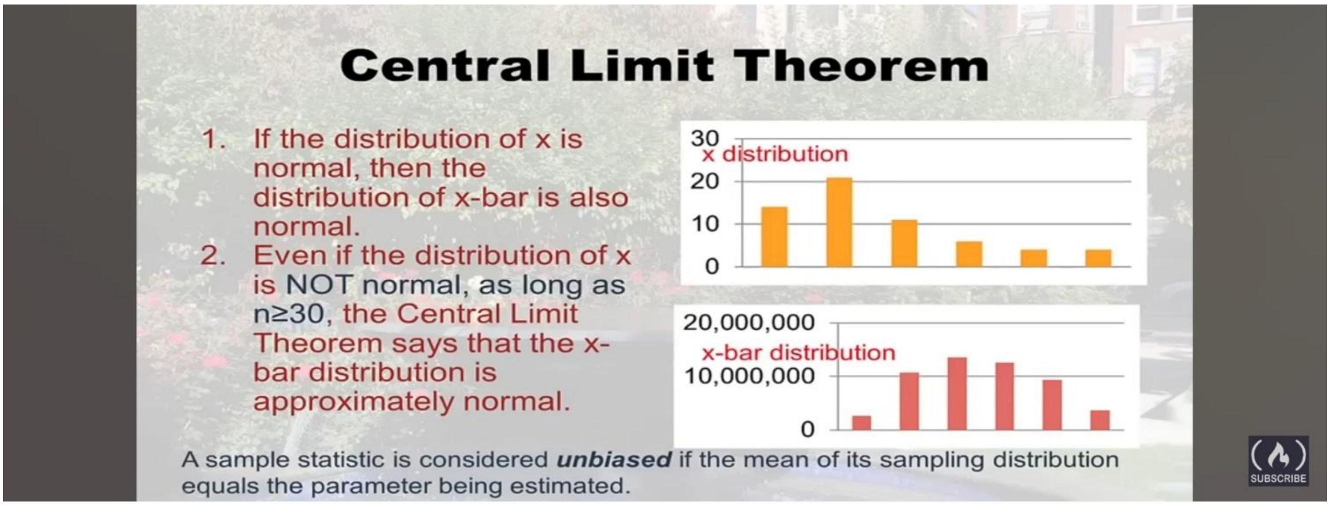
$\sigma_{\text{x-bars}}=\sigma/\sqrt{n}$

$$z=\frac{\text{x-bar} - \mu_{\text{x-bars}}}{\sigma_{\text{x-bars}}} = \frac{\text{x-bar} - \mu}{\boxed{\sigma/\sqrt{n}}}$$

....where

- n is the sample size
  - Note: Only works when n≥30!
- μ is the mean of the x distribution (population mean), and
- σ is the standard deviation of the x distribution (population standard deviation)

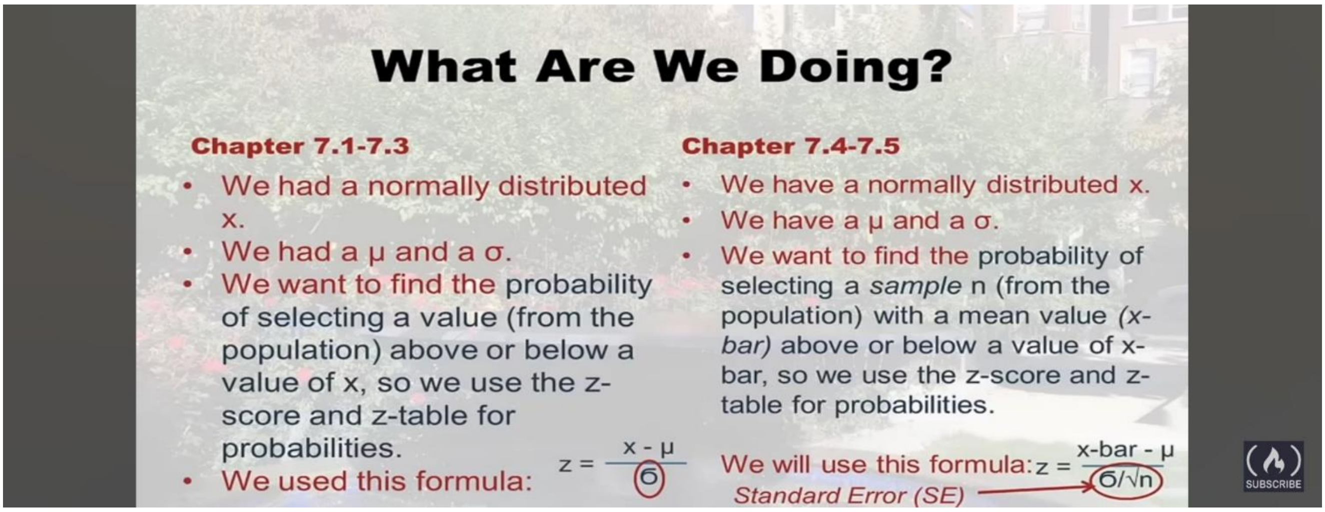
The standard error is the standard deviation of the sampling distribution.
For the x-bar sampling distribution, standard error (SE) = σ/√n



Central Limit Theorem

1. If the distribution of x is normal, then the distribution of x-bar is also normal.
2. Even if the distribution of x is NOT normal, as long as n≥30, the Central Limit Theorem says that the x-bar distribution is approximately normal.

A sample statistic is considered ***unbiased*** if the mean of its sampling distribution equals the parameter being estimated.



What Are We Doing?

Chapter 7.1-7.3

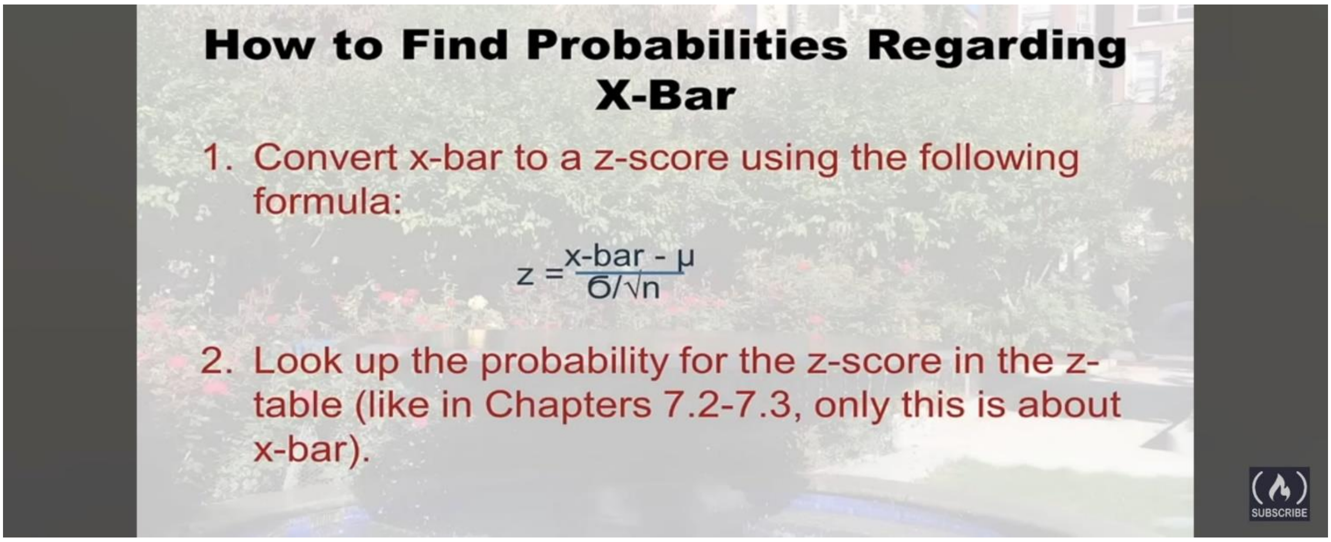
- We had a normally distributed x.
- We had a μ and a σ.
- We want to find the probability of selecting a value (from the population) above or below a value of x, so we use the z-score and z-table for probabilities.
- We used this formula: $z = \frac{x - \mu}{\sigma}$

Chapter 7.4-7.5

- We have a normally distributed x.
- We have a μ and a σ.
- We want to find the probability of selecting a *sample* n (from the population) with a mean value *(x-bar)* above or below a value of x-bar, so we use the z-score and z-table for probabilities.

We will use this formula: $z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$

Standard Error (SE)



How to Find Probabilities Regarding X-Bar

1. Convert x-bar to a z-score using the following formula:

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

2. Look up the probability for the z-score in the z-table (like in Chapters 7.2-7.3, only this is about x-bar).

Thank you